(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116011468 A
(43)申请公布日 2023.04.25

(21)申请号 202111222439.5
(22)申请日 2021.10.20
(71)申请人 珠海金山办公软件有限公司
地址 519015 广东省珠海市高新区唐家湾镇前岛环路321号金山软件园5号楼
申请人 北京金山办公软件股份有限公司
武汉金山办公软件有限公司
(72)发明人 汪保玉 王浪
(74)专利代理机构 北京路浩知识产权代理有限公司 11002
专利代理师 王宇杨
(51)Int.Cl.
G06F 40/58(2020.01)
G06N 3/0464(2023.01)
G06N 3/08(2023.01)
G06N 5/04(2023.01)
G06F 18/25(2023.01)

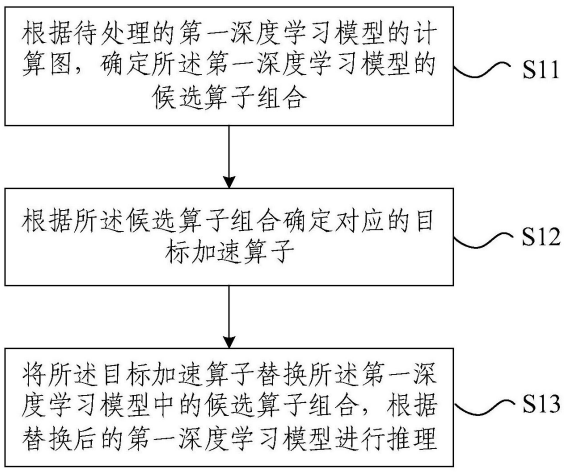
权利要求书4页 说明书35页 附图9页

(54)发明名称

深度学习模型的推理方法、机器翻译方法及装置

(57)摘要

本发明涉及一种深度学习模型的推理方法、机器翻译方法及装置。该推理方法包括:根据待处理的第一深度学习模型的计算图,确定所述第一深度学习模型的候选算子组合;根据所述候选算子组合确定对应的目标加速算子;将所述目标加速算子替换所述第一深度学习模型中的候选算子组合,根据替换后的第一深度学习模型进行推理。本发明在算子层面精确、灵活地确定了可融合的算子;根据候选算子组合确定了第一深度学习模型的加速算子;通过将候选算子组合替换为加速算子,提升了第一深度学习模型的推理效率。



CN 116011468 A

1.一种深度学习模型的推理方法,其特征在于,包括:

根据待处理的第一深度学习模型的计算图,确定所述第一深度学习模型的候选算子组合;其中,所述候选算子组合包括第一候选算子组合和/或第二候选算子组合;所述第一候选算子组合是所述第一深度学习模型的计算图中,同一分支内的具有依赖关系的多个算子的组合;所述第二候选算子组合是所述第一深度学习模型的计算图中,并行分支内具有相同上游节点的多个算子的组合;

根据所述候选算子组合确定对应的目标加速算子;其中,所述目标加速算子包括第一加速算子和/或第二加速算子;所述第一加速算子是作为样本的第二深度学习模型的计算图中,同一分支内的具有依赖关系的多个算子融合后得到的;所述第二加速算子是所述第二深度学习模型的计算图中,并行分支内具有相同上游节点的多个算子融合后得到的;

将所述目标加速算子替换所述第一深度学习模型中的候选算子组合,并根据替换后的第一深度学习模型进行推理。

2.根据权利要求1所述的深度学习模型的推理方法,其特征在于,所述根据所述候选算子组合确定对应的目标加速算子,包括:

根据所述候选算子组合在预先建立的加速算子库中查找对应的目标加速算子;其中,所述加速算子库包括算子组合与加速算子之间的映射关系;

或,

将所述候选算子组合中的算子进行融合,得到对应的目标加速算子。

3.根据权利要求1所述的深度学习模型的推理方法,其特征在于,所述将所述目标加速算子替换所述第一深度学习模型中的候选算子组合,并根据替换后的第一深度学习模型进行推理,包括:

将所述第一深度学习模型中的候选算子组合替换为所述目标加速算子;

根据替换后的第一深度学习模型进行推理测试,根据所述推理测试的结果对所述目标加速算子进行验证;

在所述目标加速算子验证合格的情况下,根据替换后的第一深度学习模型进行推理;

在所述目标加速算子验证不合格的情况下,对所述目标加速算子进行分析,根据分析结果调整所述目标加速算子,然后将所述第一深度学习模型中的候选算子组合替换为调整后的目标加速算子,并重新执行所述根据替换后的第一深度学习模型进行推理测试的步骤。

4.根据权利要求3所述的深度学习模型的推理方法,其特征在于,所述对所述目标加速算子进行分析,根据分析结果调整所述目标加速算子,包括:

在所述目标加速算子是所述第一加速算子的情况下,根据算子数据处理的先后顺序,逐步融合所述第一候选算子组合中的算子,或,在所述目标加速算子是所述第二加速算子的情况下,根据预设顺序逐步融合所述第二候选算子组合中的算子;

依次验证融合后的加速算子,确定出导致融合后的加速算子验证不合格的异常算子;

重新构建所述异常算子,并验证融合重新构建的异常算子后的加速算子,在验证合格的情况下,继续融合下一算子,直至最终融合后的所述目标加速算子验证合格。

5.根据权利要求2所述的深度学习模型的推理方法,其特征在于,在所述根据所述候选算子组合在预先建立的加速算子库中查找对应的目标加速算子之后,在所述将所述目标加

速算子替换所述第一深度学习模型中的候选算子组合之前，方法还包括：

在未为所述候选算子组合查找到对应目标加速算子的情况下，针对所述候选算子组合中的每个算子执行以下处理：在预先建立的基础算子库中查找对应所述算子的已验证算子；

将与所述候选算子组合中的多个所述算子一一对应的多个所述已验证算子进行融合，得到与所述候选算子组合所对应的目标加速算子。

6.根据权利要求5所述的深度学习模型的推理方法，其特征在于，当在预先建立的基础算子库中未查找到对应所述算子的已验证算子时，方法还包括：

创建对应所述算子的待验证算子，并对所述待验证算子进行验证，将验证合格的算子确定为所述算子对应的已验证算子。

7.根据权利要求1所述的深度学习模型的推理方法，其特征在于，所述根据待处理的第一深度学习模型的计算图，确定所述第一深度学习模型的候选算子组合，包括：

在所述第一深度学习模型的计算图的第一分支中，当至少两个相邻的算子满足第一条件、第二条件与第三条件中的任意一项时，将所述至少两个相邻的算子组成所述第一深度学习模型的第一候选算子组合；其中，所述第一条件包括：至少两个相邻的算子均为单射函数算子；所述第二条件包括：至少两个相邻的算子包括约简算子以及作为所述约简算子输入的单射函数算子；所述第三条件包括：至少两个相邻的算子包括能够融合输出的算子与逐元素复用的算子；所述第一分支为所述第一深度学习模型的计算图中的任意一个分支；

在所述第一深度学习模型的计算图的至少两个并行分支中，当存在具有相同上游节点的多个算子时，将所述多个算子组成所述第一深度学习模型的第二候选算子组合。

8.根据权利要求1至7任一项所述的深度学习模型的推理方法，其特征在于，在所述根据所述候选算子组合确定对应的目标加速算子之前，方法还包括：

根据作为样本的第二深度学习模型的计算图，确定所述第二深度学习模型的样本算子组合；其中，所述样本算子组合包括第一样本算子组合和/或第二样本算子组合；所述第一样本算子组合是所述第二深度学习模型的计算图中，同一分支内的具有依赖关系的多个样本算子的组合；所述第二样本算子组合是所述第二深度学习模型的计算图中，并行分支内具有相同上游节点的多个样本算子的组合；

获取所述样本算子组合中的各个样本算子分别对应的已验证样本算子，并保存在基础算子库中；

对所述样本算子组合中的各个样本算子所对应的已验证样本算子进行融合，得到与所述样本算子组合所对应的加速算子，并保存在加速算子库中。

9.根据权利要求8所述的深度学习模型的推理方法，其特征在于，所述获取所述样本算子组合中的各个样本算子分别对应的已验证样本算子，包括：

为所述样本算子组合中的各个样本算子分别创建对应的待验证样本算子；

对所述待验证样本算子进行验证；

验证合格后得到所述样本算子组合中的各个样本算子所对应的已验证样本算子。

10.根据权利要求9所述的深度学习模型的推理方法，其特征在于，所述对所述待验证样本算子进行验证，包括：

从标准深度学习框架中调用与所述待验证样本算子相对应的对比算子；

为所述待验证样本算子与所述对比算子设置相同的数据输入；

获取所述待验证样本算子对所述数据输入的第一推理耗时及第一推理结果；

获取所述对比算子对所述数据输入的第二推理耗时及第二推理结果；

在所述第一推理耗时小于所述第二推理耗时、且所述第一推理结果与所述第二推理结果的最大误差小于第一预设阈值的情况下，确定所述待验证样本算子验证合格。

11.根据权利要求8所述的深度学习模型的推理方法，其特征在于，所述对所述样本算子组合中的各个样本算子所对应的已验证样本算子进行融合，得到与所述样本算子组合所对应的加速算子，包括：

为所述样本算子组合中的各个样本算子分别调用对应的已验证样本算子，并对所述已验证样本算子进行融合，得到与所述样本算子组合所对应的待验证加速算子；

对所述待验证加速算子进行验证；

验证合格后得到与所述样本算子组合所对应的加速算子。

12.根据权利要求11所述的深度学习模型的推理方法，其特征在于，所述对所述待验证加速算子进行验证，包括：

为所述样本算子组合与所述待验证加速算子设置相同的数据输入；

获取所述待验证加速算子对所述数据输入的第三推理耗时及第三推理结果；

获取所述样本算子组合对所述数据输入的第四推理耗时及第四推理结果；

在所述第三推理耗时小于所述第四推理耗时、且所述第三推理结果与所述第四推理结果的最大误差小于第二预设阈值的情况下，确定所述待验证加速算子验证合格。

13.一种机器翻译方法，其特征在于，包括：

将待翻译文本输入第一机器翻译模型，得到翻译后的文本；

其中，所述第一机器翻译模型是利用目标加速算子替换初始机器翻译模型中相应的候选算子组合后得到的；所述第一机器翻译模型的获取过程包括：

根据所述初始机器翻译模型的计算图，确定所述初始机器翻译模型的候选算子组合；其中，所述候选算子组合包括第一候选算子组合和/或第二候选算子组合；所述第一候选算子组合是所述初始机器翻译模型的计算图中，同一分支内的具有依赖关系的多个算子的组合；所述第二候选算子组合是所述初始机器翻译模型的计算图中，并行分支内具有相同上游节点的多个算子的组合；

根据所述候选算子组合确定对应的目标加速算子；其中，所述目标加速算子包括第一加速算子和/或第二加速算子；所述第一加速算子是作为样本的第二机器翻译模型的计算图中，同一分支内的具有依赖关系的多个算子融合后得到的；所述第二加速算子是所述第二机器翻译模型的计算图中，并行分支内具有相同上游节点的多个算子融合后得到的；

将所述目标加速算子替换所述初始机器翻译模型中的候选算子组合，得到所述第一机器翻译模型。

14.一种深度学习模型的推理装置，其特征在于，包括：

候选算子确定模块，用于根据待处理的第一深度学习模型的计算图，确定所述第一深度学习模型的候选算子组合；其中，所述候选算子组合包括第一候选算子组合和/或第二候选算子组合；所述第一候选算子组合是所述第一深度学习模型的计算图中，同一分支内的具有依赖关系的多个算子的组合；所述第二候选算子组合是所述第一深度学习模型的计算

图中,并行分支内具有相同上游节点的多个算子的组合;

加速算子获取模块,用于根据所述候选算子组合确定对应的目标加速算子;其中,所述目标加速算子包括第一加速算子和/或第二加速算子;所述第一加速算子是作为样本的第二深度学习模型的计算图中,同一分支内的具有依赖关系的多个算子融合后得到的;所述第二加速算子是所述第二深度学习模型的计算图中,并行分支内具有相同上游节点的多个算子融合后得到的;

加速算子替换模块,用于在为所述候选算子组合查找到对应目标加速算子的情况下,将所述目标加速算子替换所述第一深度学习模型中的候选算子组合,根据替换后的第一深度学习模型进行推理。

15.一种机器翻译推理装置,其特征在于,包括:

翻译模块,用于将待翻译文本输入第一机器翻译模型,得到翻译后的文本;

其中,所述第一机器翻译模型是利用目标加速算子替换初始机器翻译模型中相应的候选算子组合后得到的;所述机器翻译推理装置,还包括:

候选算子确定模块,用于根据所述初始机器翻译模型的计算图,确定所述初始机器翻译模型的候选算子组合;其中,所述候选算子组合包括第一候选算子组合和/或第二候选算子组合;所述第一候选算子组合是所述初始机器翻译模型的计算图中,同一分支内的具有依赖关系的多个算子的组合;所述第二候选算子组合是所述初始机器翻译模型的计算图中,并行分支内具有相同上游节点的多个算子的组合;

加速算子获取模块,用于根据所述候选算子组合确定对应的目标加速算子;其中,所述加速算子包括第一加速算子和/或第二加速算子;所述第一加速算子是作为样本的第二机器翻译模型的计算图中,同一分支内的具有依赖关系的多个算子融合后得到的;所述第二加速算子是所述第二机器翻译模型的计算图中,并行分支内具有相同上游节点的多个算子融合后得到的;

加速算子替换模块,用于将所述目标加速算子替换所述初始机器翻译模型中的候选算子组合,得到所述第一机器翻译模型。

16.一种电子设备,包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序,其特征在于,所述处理器执行所述程序时实现如权利要求1至12任一项所述深度学习模型的推理方法或权利要求13所述机器翻译方法的全部或部分步骤。

17.一种计算机可读存储介质,其上存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现如权利要求1至12任一项所述深度学习模型的推理方法或权利要求13所述机器翻译方法的全部或部分步骤。

深度学习模型的推理方法、机器翻译方法及装置

技术领域

[0001] 本发明涉及人工智能技术领域，尤其涉及一种深度学习模型的推理方法、机器翻译方法及装置。

背景技术

[0002] 深度神经网络在人工智能领域中应用广泛，学术界及工业界往往使用开源机器学习框架(如Tensorflow、Pytorch等)来实现特定结构的神经网络。开源机器学习框架为了给用户较大的深度神经网络设计自由度，在底层实现了诸多细粒度的算子来满足多变的使用需求，但同时导致了推理设备在推理过程中对细粒度算子的频繁调用、数据频繁地读写及拷贝，限制了模型的推理性能。相应地影响了用户体验，也增加了推理设备的成本。

发明内容

[0003] 本发明的目的是提供一种深度学习模型的推理方法、机器翻译方法及装置。以解决现有开源推理加速框架设计出的推理模型的推理效果较差的缺陷，实现灵活高效地对模型进行推理加速，提升用户体验，降低推理设备成本。

[0004] 本发明提供一种深度学习模型的推理方法，包括：

[0005] 根据待处理的第一深度学习模型的计算图，确定所述第一深度学习模型的候选算子组合；其中，所述候选算子组合包括第一候选算子组合和/或第二候选算子组合；所述第一候选算子组合是所述第一深度学习模型的计算图中，同一分支内的具有依赖关系的多个算子的组合；所述第二候选算子组合是所述第一深度学习模型的计算图中，并行分支内具有相同上游节点的多个算子的组合；

[0006] 根据所述候选算子组合确定对应的目标加速算子；其中，所述目标加速算子包括第一加速算子和/或第二加速算子；所述第一加速算子是作为样本的第二深度学习模型的计算图中，同一分支内的具有依赖关系的多个算子融合后得到的；所述第二加速算子是所述第二深度学习模型的计算图中，并行分支内具有相同上游节点的多个算子融合后得到的；

[0007] 将所述目标加速算子替换所述第一深度学习模型中的候选算子组合，并根据替换后的第一深度学习模型进行推理。

[0008] 根据本发明提供的深度学习模型的推理方法，所述根据所述候选算子组合确定对应的目标加速算子，包括：

[0009] 根据所述候选算子组合在预先建立的加速算子库中查找对应的目标加速算子；其中，所述加速算子库包括算子组合与加速算子之间的映射关系；

[0010] 或，

[0011] 将所述候选算子组合中的算子进行融合，得到对应的目标加速算子。

[0012] 根据本发明提供的深度学习模型的推理方法，所述将所述目标加速算子替换所述第一深度学习模型中的候选算子组合，并根据替换后的第一深度学习模型进行推理，包括：

[0013] 将所述第一深度学习模型中的候选算子组合替换为所述目标加速算子；

[0014] 根据替换后的第一深度学习模型进行推理测试，根据所述推理测试的结果对所述目标加速算子进行验证；

[0015] 在所述目标加速算子验证合格的情况下，根据替换后的第一深度学习模型进行推理；

[0016] 在所述目标加速算子验证不合格的情况下，对所述目标加速算子进行分析，根据分析结果调整所述目标加速算子，然后将所述第一深度学习模型中的候选算子组合替换为调整后的目标加速算子，并重新执行所述根据替换后的第一深度学习模型进行推理测试的步骤。

[0017] 根据本发明提供的深度学习模型的推理方法，所述对所述目标加速算子进行分析，根据分析结果调整所述目标加速算子，包括：

[0018] 当所述目标加速算子是所述第一加速算子时，根据算子数据处理的先后顺序，逐步融合所述第一候选算子组合中的算子，或，当所述目标加速算子是所述第二加速算子时，根据预设顺序逐步融合所述第二候选算子组合中的算子；

[0019] 依次验证融合后的加速算子，确定出导致融合后的加速算子验证不合格的异常算子；

[0020] 重新构建所述异常算子，并验证融合重新构建的异常算子后的加速算子，在验证合格的情况下，继续融合下一算子，直至最终融合后的所述目标加速算子验证合格。

[0021] 根据本发明提供的深度学习模型的推理方法，在所述根据所述候选算子组合在预先建立的加速算子库中查找对应的目标加速算子之后，在所述将所述目标加速算子替换所述第一深度学习模型中的候选算子组合之前，方法还包括：

[0022] 在未为所述候选算子组合查找到对应目标加速算子的情况下，针对所述候选算子组合中的每个算子执行以下处理：在预先建立的基础算子库中查找对应所述算子的已验证算子；

[0023] 将与所述候选算子组合中的多个所述算子一一对应的多个所述已验证算子进行融合，得到与所述候选算子组合所对应的目标加速算子。

[0024] 根据本发明提供的深度学习模型的推理方法，当在预先建立的基础算子库中未查找到对应所述算子的已验证算子时，方法还包括：

[0025] 创建对应所述算子的待验证算子，并对所述待验证算子进行验证，将验证合格的算子确定为所述算子对应的已验证算子。

[0026] 根据本发明提供的深度学习模型的推理方法，所述根据待处理的第一深度学习模型的计算图，确定所述第一深度学习模型的候选算子组合，包括：

[0027] 在所述第一深度学习模型的计算图的第一分支中，当至少两个相邻的算子满足第一条件、第二条件与第三条件中的任意一项时，将所述至少两个相邻的算子组成所述第一深度学习模型的第一候选算子组合；其中，所述第一条件包括：至少两个相邻的算子均为单射函数算子；所述第二条件包括：至少两个相邻的算子包括约简算子以及作为所述约简算子输入的单射函数算子；所述第三条件包括：至少两个相邻的算子包括能够融合输出的算子与逐元素复用的算子；所述第一分支为所述第一深度学习模型的计算图中的任意一个分支；

[0028] 在所述第一深度学习模型的计算图的至少两个并行分支中，当存在具有相同上游节点的多个算子时，将所述多个算子组成所述第一深度学习模型的第二候选算子组合。

[0029] 根据本发明提供的深度学习模型的推理方法，在所述根据所述候选算子组合确定对应的目标加速算子之前，方法还包括：

[0030] 根据作为样本的第二深度学习模型的计算图，确定所述第二深度学习模型的样本算子组合；其中，所述样本算子组合包括第一样本算子组合和/或第二样本算子组合；所述第一样本算子组合是所述第二深度学习模型的计算图中，同一分支内的具有依赖关系的多个样本算子的组合；所述第二样本算子组合是所述第二深度学习模型的计算图中，并行分支内具有相同上游节点的多个样本算子的组合；

[0031] 获取所述样本算子组合中的各个样本算子分别对应的已验证样本算子，并保存在基础算子库中；

[0032] 对所述样本算子组合中的各个样本算子所对应的已验证样本算子进行融合，得到与所述样本算子组合所对应的加速算子，并保存在加速算子库中。

[0033] 根据本发明提供的深度学习模型的推理方法，所述获取所述样本算子组合中的各个样本算子分别对应的已验证样本算子，包括：

[0034] 为所述样本算子组合中的各个样本算子分别创建对应的待验证样本算子；

[0035] 对所述待验证样本算子进行验证；

[0036] 验证合格后得到所述样本算子组合中的各个样本算子所对应的已验证样本算子。

[0037] 根据本发明提供的深度学习模型的推理方法，所述对所述待验证样本算子进行验证，包括：

[0038] 从标准深度学习框架中调用与所述待验证样本算子相对应的对比算子；

[0039] 为所述待验证样本算子与所述对比算子设置相同的数据输入；

[0040] 获取所述待验证样本算子对所述数据输入的第一推理耗时及第一推理结果；

[0041] 获取所述对比算子对所述数据输入的第二推理耗时及第二推理结果；

[0042] 在所述第一推理耗时小于所述第二推理耗时、且所述第一推理结果与所述第二推理结果的最大误差小于第一预设阈值的情况下，确定所述待验证样本算子验证合格。

[0043] 根据本发明提供的深度学习模型的推理方法，所述对所述样本算子组合中的各个样本算子所对应的已验证样本算子进行融合，得到与所述样本算子组合所对应的加速算子，包括：

[0044] 为所述样本算子组合中的各个样本算子分别调用对应的已验证样本算子，并对所述已验证样本算子进行融合，得到与所述样本算子组合所对应的待验证加速算子；

[0045] 对所述待验证加速算子进行验证；

[0046] 验证合格后得到与所述样本算子组合所对应的加速算子。

[0047] 根据本发明提供的深度学习模型的推理方法，所述对所述待验证加速算子进行验证，包括：

[0048] 为所述样本算子组合与所述待验证加速算子设置相同的数据输入；

[0049] 获取所述待验证加速算子对所述数据输入的第三推理耗时及第三推理结果；

[0050] 获取所述样本算子组合对所述数据输入的第四推理耗时及第四推理结果；

[0051] 在所述第三推理耗时小于所述第四推理耗时、且所述第三推理结果与所述第四推

理结果的最大误差小于第二预设阈值的情况下，确定所述待验证加速算子验证合格。

[0052] 本发明还提供一种机器翻译方法，包括：

[0053] 将待翻译文本输入第一机器翻译模型，得到翻译后的文本；

[0054] 其中，所述第一机器翻译模型是利用目标加速算子替换初始机器翻译模型中相应的候选算子组合后得到的；所述第一机器翻译模型的获取过程包括：

[0055] 根据所述初始机器翻译模型的计算图，确定所述初始机器翻译模型的候选算子组合；其中，所述候选算子组合包括第一候选算子组合和/或第二候选算子组合；所述第一候选算子组合是所述初始机器翻译模型的计算图中，同一分支内的具有依赖关系的多个算子的组合；所述第二候选算子组合是所述初始机器翻译模型的计算图中，并行分支内具有相同上游节点的多个算子的组合；

[0056] 根据所述候选算子组合确定对应的目标加速算子；其中，所述目标加速算子包括第一加速算子和/或第二加速算子；所述第一加速算子是作为样本的第二机器翻译模型的计算图中，同一分支内的具有依赖关系的多个算子融合后得到的；所述第二加速算子是所述第二机器翻译模型的计算图中，并行分支内具有相同上游节点的多个算子融合后得到的；

[0057] 将所述目标加速算子替换所述初始机器翻译模型中的候选算子组合，得到所述第一机器翻译模型。

[0058] 根据本发明提供的机器翻译方法，所述根据所述候选算子组合确定对应的目标加速算子，包括：

[0059] 根据所述候选算子组合在预先建立的加速算子库中查找对应的目标加速算子；其中，所述加速算子库包括算子组合与加速算子之间的映射关系；

[0060] 或，

[0061] 将所述候选算子组合中的算子进行融合，得到对应的目标加速算子。

[0062] 根据本发明提供的机器翻译方法，所述将所述目标加速算子替换所述初始机器翻译模型中的候选算子组合，得到所述第一机器翻译模型，包括：

[0063] 将所述初始机器翻译模型中的候选算子组合替换为所述目标加速算子；

[0064] 根据替换后的初始机器翻译模型进行推理测试，根据所述推理测试的结果对所述目标加速算子验证；

[0065] 在所述目标加速算子验证合格的情况下，将替换后的初始机器翻译模型确定为所述第一机器翻译模型；

[0066] 在所述目标加速算子验证不合格的情况下，对所述目标加速算子进行分析，根据分析结果调整所述目标加速算子，然后将所述初始机器翻译模型中的候选算子组合替换为调整后的目标加速算子，并重新执行所述根据替换后的初始机器翻译模型进行推理测试的步骤。

[0067] 根据本发明提供的机器翻译方法，所述对所述目标加速算子进行分析，根据分析结果调整所述目标加速算子，包括：

[0068] 当所述目标加速算子是所述第一加速算子时根据算子数据处理的先后顺序，逐步融合所述第一候选算子组合中的算子，或，当所述目标加速算子是所述第二加速算子时，根据预设顺序逐步融合所述第二候选算子组合中的算子；

[0069] 依次验证融合后的加速算子，确定出导致融合后的加速算子验证不合格的异常算子；

[0070] 重新构建所述异常算子，并验证融合重新构建的异常算子后的加速算子，在验证合格的情况下，继续融合下一算子，直至最终融合后的所述目标加速算子验证合格。

[0071] 根据本发明提供的机器翻译方法，在所述根据所述候选算子组合在预先建立的加速算子库中查找对应的目标加速算子之后，在所述将所述目标加速算子替换所述初始机器翻译模型中的候选算子组合之前，方法还包括：

[0072] 在未为所述候选算子组合查找到对应目标加速算子的情况下，针对所述候选算子组合中的每个算子执行以下处理：在预先建立的基础算子库中查找对应所述算子的已验证算子；

[0073] 将与所述候选算子组合中的多个所述算子一一对应的多个所述已验证算子进行融合，得到与所述候选算子组合所对应的目标加速算子。

[0074] 根据本发明提供的机器翻译方法，当在预先建立的基础算子库中未查找到对应所述算子的已验证算子时，方法还包括：

[0075] 创建对应所述算子的待验证算子，并对所述待验证算子进行验证，将验证合格的算子确定为所述算子对应的已验证算子。

[0076] 根据本发明提供的机器翻译方法，所述根据所述初始机器翻译模型的计算图，确定所述第一机器翻译模型的候选算子组合，包括：

[0077] 在所述初始机器翻译模型的计算图的第一分支中，当至少两个相邻的算子满足第一条件、第二条件与第三条件中的任意一项时，将所述至少两个相邻的算子组成所述初始机器翻译模型的第一候选算子组合；其中，所述第一条件包括：至少两个相邻的算子均为单射函数算子；所述第二条件包括：至少两个相邻的算子包括约简算子以及作为所述约简算子输入的单射函数算子；所述第三条件包括：至少两个相邻的算子包括能够融合输出的算子与逐元素复用的算子；所述第一分支为所述初始机器翻译模型的计算图中的任意一个分支；

[0078] 在所述初始机器翻译模型的计算图的至少两个并行分支中，当存在具有相同上游节点的多个算子时，将所述多个算子组成所述初始机器翻译模型的第二候选算子组合。

[0079] 根据本发明提供的机器翻译方法，在所述根据所述候选算子组合确定对应的目标加速算子之前，方法还包括：

[0080] 根据作为样本的第二机器翻译模型的计算图，确定所述第二机器翻译模型的样本算子组合；其中，所述样本算子组合包括第一样本算子组合和/或第二样本算子组合；所述第一样本算子组合是所述第二机器翻译模型的计算图中，同一分支内的具有依赖关系的多个样本算子的组合；所述第二样本算子组合是所述第二机器翻译模型的计算图中，并行分支内具有相同上游节点的多个样本算子的组合；

[0081] 获取所述样本算子组合中的各个样本算子分别对应的已验证样本算子，并保存在基础算子库中；

[0082] 对所述样本算子组合中的各个样本算子所对应的已验证样本算子进行融合，得到与所述样本算子组合所对应的加速算子，并保存在加速算子库中。

[0083] 根据本发明提供的机器翻译方法，所述获取所述样本算子组合中的各个样本算子

分别对应的已验证样本算子，包括：

[0084] 为所述样本算子组合中的各个样本算子分别创建对应的待验证样本算子；

[0085] 对所述待验证样本算子进行验证；

[0086] 验证合格后得到所述样本算子组合中的各个样本算子所对应的已验证样本算子。

[0087] 根据本发明提供的机器翻译方法，所述对所述待验证样本算子进行验证，包括：

[0088] 从标准深度学习框架中调用与所述待验证样本算子相对应的对比算子；

[0089] 为所述待验证样本算子与所述对比算子设置相同的数据输入；

[0090] 获取所述待验证样本算子对所述数据输入的第一推理耗时及第一推理结果；

[0091] 获取所述对比算子对所述数据输入的第二推理耗时及第二推理结果；

[0092] 在所述第一推理耗时小于所述第二推理耗时、且所述第一推理结果与所述第二推理结果的最大误差小于第一预设阈值的情况下，确定所述待验证样本算子验证合格。

[0093] 根据本发明提供的机器翻译方法，所述对所述样本算子组合中的各个样本算子所对应的已验证样本算子进行融合，得到与所述样本算子组合所对应的加速算子，包括：

[0094] 为所述样本算子组合中的各个样本算子分别调用对应的已验证样本算子，并对所述已验证样本算子进行融合，得到与所述样本算子组合所对应的待验证加速算子；

[0095] 对所述待验证加速算子进行验证；

[0096] 验证合格后得到与所述样本算子组合所对应的加速算子。

[0097] 根据本发明提供的机器翻译方法，所述对所述待验证加速算子进行性能和准确度的验证，包括：

[0098] 为所述样本算子组合与所述待验证加速算子设置相同的数据输入；

[0099] 获取所述待验证加速算子对所述数据输入的第三推理耗时及第三推理结果；

[0100] 获取所述样本算子组合对所述数据输入的第四推理耗时及第四推理结果；

[0101] 在所述第三推理耗时小于所述第四推理耗时、且所述第三推理结果与所述第四推理结果的最大误差小于第二预设阈值的情况下，确定所述待验证加速算子验证合格。

[0102] 本发明还提供一种深度学习模型的推理装置，包括：

[0103] 候选算子确定模块，用于根据待处理的第一深度学习模型的计算图，确定所述第一深度学习模型的候选算子组合；其中，所述候选算子组合包括第一候选算子组合和/或第二候选算子组合；所述第一候选算子组合是所述第一深度学习模型的计算图中，同一分支内的具有依赖关系的多个算子的组合；所述第二候选算子组合是所述第一深度学习模型的计算图中，并行分支内具有相同上游节点的多个算子的组合；

[0104] 加速算子获取模块，用于根据所述候选算子组合确定对应的目标加速算子；其中，所述目标加速算子包括第一加速算子和/或第二加速算子；所述第一加速算子是作为样本的第二深度学习模型的计算图中，同一分支内的具有依赖关系的多个算子融合后得到的；所述第二加速算子是所述第二深度学习模型的计算图中，并行分支内具有相同上游节点的多个算子融合后得到的；

[0105] 加速算子替换模块，用于在为所述候选算子组合查找到对应目标加速算子的情况下，将所述目标加速算子替换所述第一深度学习模型中的候选算子组合，根据替换后的第一深度学习模型进行推理。

[0106] 本发明还提供一种机器翻译推理装置，包括：

[0107] 翻译模块，用于将待翻译文本输入第一机器翻译模型，得到翻译后的文本；

[0108] 其中，所述第一机器翻译模型是利用目标加速算子替换初始机器翻译模型中相应的候选算子组合后得到的；所述机器翻译推理装置，还包括：

[0109] 候选算子确定模块，用于根据所述初始机器翻译模型的计算图，确定所述初始机器翻译模型的候选算子组合；其中，所述候选算子组合包括第一候选算子组合和/或第二候选算子组合；所述第一候选算子组合是所述初始机器翻译模型的计算图中，同一分支内的具有依赖关系的多个算子的组合；所述第二候选算子组合是所述初始机器翻译模型的计算图中，并行分支内具有相同上游节点的多个算子的组合；

[0110] 加速算子获取模块，用于根据所述候选算子组合确定对应的目标加速算子；其中所述目标加速算子包括第一加速算子和/或第二加速算子；所述第一加速算子是作为样本的第二机器翻译模型的计算图中，同一分支内的具有依赖关系的多个算子融合后得到的；所述第二加速算子是所述第二机器翻译模型的计算图中，并行分支内具有相同上游节点的多个算子融合后得到的；

[0111] 加速算子替换模块，用于将所述目标加速算子替换所述初始机器翻译模型中的候选算子组合，得到所述第一机器翻译模型。

[0112] 本发明还提供一种电子设备，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，所述处理器执行所述程序时实现如上任一项所述深度学习模型的推理方法或所述机器翻译方法的全部或部分步骤。

[0113] 本发明还提供一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现如上任一项所述深度学习模型的推理方法或所述机器翻译方法的全部或部分步骤。

[0114] 本发明提供的一种深度学习模型的推理方法、机器翻译方法、装置、电子设备、存储介质及程序，根据待处理的第一深度学习模型的计算图中同一分支内的具有依赖关系的算子或并行分支内具有相同上游节点的算子确定候选算子组合，在算子层面精确、灵活地确定了可融合的算子；根据候选算子组合确定了第一深度学习模型的加速算子；通过将候选算子组合替换为加速算子，提升了第一深度学习模型的推理效率，实现灵活高效地对模型进行推理加速，提升了用户体验，降低了推理设备成本。

附图说明

[0115] 图1是本发明提供的一种深度学习模型的推理方法的流程示意图；

[0116] 图2是本发明示例的待处理的第一深度学习模型的计算图之一；

[0117] 图3是本发明示例的待处理的第一深度学习模型的计算图之二；

[0118] 图4是本发明提供的一种深度学习模型的推理方法的一个实施例中的程序流程图；

[0119] 图5是本发明提供的一种深度学习模型的推理方法的一个实施例中预先构建加速算子库的程序流程图；

[0120] 图6是本发明提供的一种深度学习模型的推理方法中算子融合前的计算图示例之一；

[0121] 图7是本发明提供的一种深度学习模型的推理方法中算子融合后的计算图示例之

二；

[0122] 图8是本发明提供的一种深度学习模型的推理方法中算子融合前的计算图示例之三；

[0123] 图9是本发明提供的一种深度学习模型的推理方法中算子融合后的计算图示例之四；

[0124] 图10是本发明提供的一种深度学习模型的推理方法中算子融合后的计算图示例之五；

[0125] 图11是本发明提供的一种机器翻译方法的流程示意图；

[0126] 图12是本发明提供的一种深度学习模型的推理装置的结构示意图；

[0127] 图13是本发明提供的一种机器翻译装置的结构示意图；

[0128] 图14是本发明提供的电子设备的结构示意图。

具体实施方式

[0129] 为使本发明的目的、技术方案和优点更加清楚，下面将结合本发明中的附图，对本发明中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0130] 下面结合图1-图14描述本发明的一种深度学习模型的推理方法、机器翻译方法、装置、电子设备、存储介质及程序

[0131] 图1是本发明提供的一种深度学习模型的推理方法的流程示意图，如图1所示，该方法包括：

[0132] S11、根据待处理的第一深度学习模型的计算图，确定第一深度学习模型的候选算子组合；其中，候选算子组合包括第一候选算子组合和/或第二候选算子组合；第一候选算子组合是第一深度学习模型的计算图中，同一分支内的具有依赖关系的多个算子的组合；第二候选算子组合是第一深度学习模型的计算图中，并行分支内具有相同上游节点的多个算子的组合。

[0133] 具体地，计算图用于将计算过程图形化。深度学习模型将输入数据经过一系列数据处理得到推理结果，深度学习模型中的对数据执行的各种处理操作即可抽象为算子，算子对应于计算图中的算子节点，可以理解的是计算图中的算子节点可以是独立整体的单个算子，也可以是由更细粒度的算子的有序组合；深度学习模型中的各步骤之间的数据流转可抽象为算子之间的依赖关系(即一个算子的输出或部分输出是另一个算子的输入或部分输入)，算子之间的依赖关系对应于计算图中的算子节点之间的连接关系。

[0134] 根据待处理的第一深度学习模型的计算图，确定出第一深度学习模型的可以进行算子融合的候选算子组合。

[0135] 候选算子组合可以是第一深度学习模型的计算图中同一分支内的具有依赖关系的多个算子组成的第一候选算子组合，其中“多个”是指“至少两个”。举例说明如下：图2是本发明示例的待处理的第一深度学习模型的计算图之一，如图2所示，算子1、算子2、算子3处于第一深度学习模型的计算图的同一分支，且依次具有依赖关系，此时第一候选算子组合可以是{算子1，算子2}，可以是{算子2，算子3}，还可以是{算子1，算子2，算子3}。同一分

支内具有依赖关系的算子之间的输入数据输出数据存在关联，将其确定为候选算子组合用于进行后续算子融合，可以有效减少第一深度学习模型对于数据的存取次数，提升模型推理效率。

[0136] 候选算子组合还可以是第一深度学习模型的计算图中并行分支内的具有相同上游节点的多个算子组成的第二候选算子组合，其中“多个”是指“至少两个”。举例说明如下：图3是本发明示例的待处理的第一深度学习模型的计算图之二，如图3所示，计算图中包括左、中、右三个并行分支，以及算子1-算子6共六种算子，此时，第二候选算子组合可以是左、中两个并行分支中的两个算子4组成的算子组合，还可以是左、中、右三个并行分支中的三个算子6组成的算子组合。并行分支内具有相同上游节点的多个算子意味着该多个算子的数据输入节点相同，将其确定为候选算子组合用于进行后续算子融合，可以有效减少第一深度学习模型对于数据的存取次数，提升模型推理效率。

[0137] 进一步地，并行分支内具有相同上游节点的多个算子可以是相同的算子，相同的算子执行的数据处理内容相同，因此，并行分支内具有相同上游节点的多个相同算子可以组成第二候选算子组合，例如图3中所示的左、中、右三个并行分支中的三个算子6可以组成第二候选算子组合。

[0138] 并行分支内具有相同上游节点的多个算子还可以是输入数据维度相同的不同算子，举例来说，第一分支的算子b_1与第二分支的算子b_2对相同的上游节点a输入的二维数据进行处理，b_1为二维卷积算子，b_2为四维卷积算子，则b_2可以拆分出一个二维卷积算子，可以将b_2拆出的二维卷积算子部分与b_1进行融合，可见，b_1，b_2也可以确定为第二候选算子组合。相同数据输入数据的维度相同才能保证融合的线程相同，进而可以进行算子融合，减少融合后算子的种类。

[0139] S12、根据候选算子组合确定对应的目标加速算子；其中，目标加速算子包括第一加速算子和/或第二加速算子；第一加速算子是作为样本的第二深度学习模型的计算图中，同一分支内的具有依赖关系的多个算子融合后得到的；第二加速算子是第二深度学习模型的计算图中，并行分支内具有相同上游节点的多个算子融合后得到的；

[0140] 具体地，融合算子可以解决模型推理或训练过程中的数据频繁读写，同时，减少中间结果的写回操作，降低访存操作。现有的推理加速框架如TensorRT等，使用条件较为局限，如果模型中含有不受支持的算子，将无法使用这些推理加速工具。并且，现有的推理加速框架只能够从整个模型进层面行优化，推理加速效果较差。本申请中候选算子组合对应的目标加速算子可以是根据作为样本的第二深度学习模型的计算图中同一分支内的具有依赖关系的多个算子融合后得到的第一加速算子，还可以是第二深度学习模型的计算图中并行分支内具有相同上游节点的多个算子融合后得到的第二加速算子。依赖关系是指一个算子的输出(或者输出的一部分)是另一个算子的输入(或者输入的一部分)，此时这两个算子存在依赖关系。算子融合的目的是减少整个计算图中算子节点的数量，对应到推理设备来说，从一个算子节点到另一个算子节点之间就意味着数据的搬运，减少算子节点相应地可以减少数据的存取调用，进而提升推理设备的工作效率。算子融合的实现方式可以是：根据待融合的算子组合中的各个算子，基于异构编程(如CPU+GPU编程)生成各个算子的核函数(即程序语言编写的实现算子的代码)，并根据各个算子的核函数生成算子组合的核函数，然后进一步利用编译器(如nvcc编译器)对算子组合的核函数进行编译即可得到算子组

合融合后的加速算子。需要说明的是，深度学习模型的训练/推理过程通常涉及大量的并行运算，基于这一特性，利用异构编程实现的深度学习模型可以部署于异构计算系统，从而大幅提升深度学习模型的训练/推理效率。异构计算系统通常由通用处理器和许多特定于域的处理器组成：通用处理器作为控制设备(称为主机)，用于复杂的控制和调度；特定于域的处理器作为子设备(称为MLU)，用于大规模并行计算和特定于域的计算任务。主机和MLU合作完成计算任务。对于异构计算系统，原始的同构并行编程模型不再适用，因此需要用到异构编程。

[0141] S13、将目标加速算子替换第一深度学习模型中的候选算子组合，根据替换后的第一深度学习模型进行推理。

[0142] 具体地，确定目标加速算子后，即可将目标加速算子替换第一深度学习模型中的候选算子组合，将替换后的第一深度学习模型用于进行推理。

[0143] 本实施例中根据待处理的第一深度学习模型的计算图中同一分支内的具有依赖关系的算子或并行分支内具有相同上游节点的算子确定候选算子组合，在算子层面精确、灵活地确定了可融合的算子；根据候选算子组合确定了第一深度学习模型的目标加速算子；通过将候选算子组合替换为加速算子，提升了第一深度学习模型的推理效率，实现灵活高效地对模型进行推理加速，提升了用户体验，降低了推理设备成本。

[0144] 基于上述实施例，在一个实施例中，根据候选算子组合确定对应的目标加速算子，包括：根据候选算子组合在预先建立的加速算子库中查找对应的目标加速算子；其中，加速算子库包括算子组合与加速算子之间的映射关系；或，将所选算子组合中的算子进行融合，得到对应的目标加速算子。

[0145] 具体地，可以基于预先建立的加速算子库中查找对应的目标加速算子，加速算子库中包括，作为样本的第二深度学习模型的计算图中同一分支内的具有依赖关系的多个算子融合后得到的第一加速算子，以及第二深度学习模型的计算图中，并行分支内具有相同上游节点的多个算子融合后得到的第二加速算子。加速算子库中还包括算子组合与加速算子之间的映射关系，根据第一深度学习模型的候选算子组合，结合映射关系即可确定候选算子组合对应的目标加速算子。通过预先收集作为样本的第二深度学习模型，丰富加速算子库中的加速算子，可以有效提升第一深度学习模型的确定加速算子的效率。

[0146] 还可以基于候选算子组合中的算子直接进行算子融合，得到对应的目标加速算子，相对于上述实施例，实时融合候选算子组合中的算子能够减少在加速算子库中的查找时间。

[0147] 还可以当未在预先建立的加速算子库中查找到对应的目标加速算子时，基于候选算子组合中的算子直接进行算子融合，得到对应的目标加速算子，从而能够确保目标加速算子的获取机率。

[0148] 本实施例中根据候选算子组合，基于预先建立的算子库或直接进行算子融合，全面准确地确定了目标加速算子，进一步提高了深度学习模型的推理效率。

[0149] 基于上述实施例，在一个实施例中，将目标加速算子替换第一深度学习模型中的候选算子组合，并根据替换后的第一深度学习模型进行推理，包括：

[0150] 将第一深度学习模型中的候选算子组合替换为目标加速算子；

[0151] 根据替换后的第一深度学习模型进行推理测试，根据推理测试的结果对目标加速

算子进行验证；

[0152] 在目标加速算子验证合格的情况下，根据替换后的第一深度学习模型进行推理；

[0153] 在目标加速算子验证不合格的情况下，对目标加速算子进行分析，并根据分析结果调整目标加速算子，然后将第一深度学习模型中的候选算子组合替换为调整后的目标加速算子，并重新执行根据替换后的第一深度学习模型进行推理测试的步骤。

[0154] 具体地，将替换后的第一深度学习模型用于进行推理之前，还需要对替换后的第一深度学习模型进行验证，验证目标加速算子的加速效果，在目标加速算子验证合格，即目标加速算子的性能满足性能预设条件且目标加速算子的准确度满足准确度预设条件的情况下才可以用于替换进行推理。推理测试的具体内容可以是：在相同推理设备条件下，对替换前后的第一深度学习模型给予相同的数据输入，记录推理结果以及推理耗时。目标加速算子的性能的预设条件例如：替换后的推理耗时的平均值小于与替换前的推理耗时的平均值；目标加速算子的准确度预设条件例如：替换后的输出结果与替换前的输出结果之间的最大误差小于预设阈值(例如1e-4，即最大误差小于10的-4次方)。在目标加速算子的性能和准确度满足预设条件的情况下，即可根据替换后的第一深度学习模型进行推理。

[0155] 在目标加速算子验证不合格，即目标加速算子的性能不满足性能预设条件或准确度不满足准确度预设条件，则说明将候选算子组合替换为目标加速算子会导致第一深度学习模型的推理速度得不到提升或者推理的准确率降低，因此，对于目标加速算子验证不合格的情况，需要对目标加速算子进行分析，根据分析结果重新调整目标加速算子，然后将第一深度学习模型中的候选算子组合替换为调整后的目标加速算子，并重新执行根据替换后的第一深度学习模型进行推理测试的步骤，直至得到验证合格的目标加速算子。

[0156] 本实施例中，在将替换后的第一深度学习模型用于进行推理之前，对替换后的第一深度学习模型进行推理测试，验证目标加速算子的加速效果，保障了第一深度学习模型的推理准确率以及推理加速效果，在目标加速算子验证不合格的情况下，对目标加速算子进行分析，根据分析结果调整加速算子，保障了替换后得到的第一深度学习模型的推理准确率，并提升了推理速度。

[0157] 基于上述任一实施例，在一个实施例中，对目标加速算子进行分析，根据分析结果调整目标加速算子，包括：

[0158] 当目标加速算子是第一加速算子时，根据算子数据处理的先后顺序，逐步融合第一候选算子组合中的算子，或，当目标加速算子是第二加速算子时，根据预设顺序逐步融合第二候选算子组合中的算子；

[0159] 依次验证融合后的加速算子，确定出导致融合后的加速算子验证不合格的异常算子；

[0160] 重新构建异常算子，并验证融合重新构建的异常算子后的加速算子，在验证合格的情况下，继续融合下一算子，直至最终融合后的目标加速算子验证合格。

[0161] 具体地，可以通过逐步融合算子的方式对目标加速算子进行分析，可以理解的是，当目标加速算子是上述第一加速算子时，此时目标加速算子对应的候选算子组合是根据同一分支内具有依赖关系的多个算子确定出的第一候选算子组合，进行分析时可以直接根据算子数据处理的先后顺序逐步融合第一候选算子组合中的算子。具体而言，对于同一分支内具有依赖关系的算子，融合算子的顺序即数据流转的先后顺序。当目标加速算子是第二

加速算子时，此时目标加速算子对应的候选算子组合是根据并行分支内具有相同上游节点的多个算子确定出的第二候选算子组合，由于数据是并行处理的，数据处理并无严格的先后处理顺序，进行分析时可以按照预设的顺序进行逐步融合算子，预设的顺序例如多个算子的标识序号由小到大的顺序。

[0162] 逐步融合算子后，依次验证融合后的加速算子，验证内容可以包括融合算子的性能和准确度，将导致融合后性能或准确度不满足预设条件的算子确定为异常算子。在确定出异常算子后，重新生成异常算子及相应的融合后的加速算子，具体地，可以基于异构编程重新生成异常算子的核函数(即程序语言编写的实现算子的代码)，对异常算子逐步排查计算结果，对有问题的计算步骤改写为正确的计算步骤，根据重新生成的异常算子的核函数以及上游算子的核函数得到算子组合的核函数，然后进一步对算子组合的核函数进行编译即可得到算子组合再次融合后的加速算子。进一步地，重新验证融合异常算子后的加速算子，在验证合格的情况下，继续融合下一算子，同理，对于下一异常算子重复上述步骤，直至最终融合后得到的目标加速算子的性能和准确度满足预设条件。

[0163] 本实施例中，通过逐步融合算子并进行验证确定出了导致目标加速算子验证不合格的异常算子，对异常算子逐步排查计算结果，对有问题的计算步骤改写为正确的计算步骤，通过重新生成异常算子以及融合异常算子后的加速算子并进行验证，对目标加速算子重新生成并验证性能及准确性，进一步保障了替换后的第一深度学习模型的推理结果与替换前的推理结果的一致性，并提升了推理速度。

[0164] 基于上述任一实施例，在一个实施例中，在根据候选算子组合确定对应的目标加速算子之后，在将目标加速算子替换第一深度学习模型中的候选算子组合之前，方法还包括：

[0165] 在未为候选算子组合查找到对应目标加速算子的情况下，针对候选算子组合中的每个算子执行以下处理：在预先建立的基础算子库中查找对应算子的已验证算子；

[0166] 将与候选算子组合中的多个算子一一对应的多个已验证算子进行融合，得到与候选算子组合所对应的目标加速算子。

[0167] 具体地，本实施例中还预先建立了基础算子库，基础算子库中包括多个可用于进行算子融合的预先经过验证的已验证算子。在无法利用加速算子库直接获取候选算子组合对应的目标加速算子的情况下，可以利用预先建立的基础算子库生成目标加速算子。若候选算子组合中的每个算子经查找确定在基础算子库都存在所对应的已验证算子，此时可以直接利用基础算子库中每个算子对应的已验证算子进行融合得到与候选算子组合所对应的目标加速算子，提高获取目标加速算子的效率。

[0168] 本实施例中，通过基础算子库，提高了获取目标加速算子的效率。

[0169] 基于上述任一实施例，在一个实施例中，当在预先建立的基础算子库中未查找到对应算子的已验证算子时，方法还包括：

[0170] 创建对应算子的待验证算子，并对待验证算子进行验证，将验证合格的算子确定为算子对应的已验证算子。

[0171] 具体地，对于候选算子组合中的算子，若在基础算子库中不存在对应于该算子的已验证算子，则需要重新创建待验证算子。具体地，可以基于异构编程生成该算子的核函数(即程序语言编写的实现算子的代码)，经过编译器编译后得到重新创建的待验证算子，并

对重新创建的待验证算子进行验证，验证内容可以包括性能和准确度。验证过程可以通过如下方式实现：为待验证样本算子与对比算子设置相同的数据输入，并记录待验证样本算子与对比算子的推理结果、推理耗时。根据推理结果验证待验证样本算子准确度是否满足准确度预设条件，根据推理耗时验证待验证样本算子的性能是否满足性能预设条件。对比算子可以是已有的深度学习框架如pytorch、tensorflow中与待验证算子对应的对比算子，性能合格的预设条件可以是待验证样本算子的推理耗时小于对比算子的推理耗时，准确度合格的预设条件可以是待验证样本算子的推理结果与对比算子的推理结果之间的最大误差小于预设阈值(例如1e-4，即10的-4次方)。在待验证算子满足预设条件的情况下，将已验证算子加入基础算子库中。重复上述步骤，直至在基础算子库中可以查找到候选算子组合中的各个算子所对应的已验证算子。

[0172] 本实施例中，在基础算子库中不存在与候选算子组合中的算子对应的已验证算子的情况下，分别创建并验证待验证算子，便于顺利地获取目标加速算子。

[0173] 基于上述任一实施例，在一个实施例中，根据待处理的第一深度学习模型的计算图，确定第一深度学习模型的候选算子组合，包括：

[0174] 在第一深度学习模型的计算图的第一分支中，当至少两个相邻的算子满足第一条件、第二条件与第三条件中的任意一项时，则将至少两个相邻的算子组成第一深度学习模型的第一候选算子组合；其中，第一条件包括：至少两个相邻的算子均为单射函数算子；第二条件包括：至少两个相邻的算子包括约简算子以及作为约简算子输入的单射函数算子；第三条件包括：至少两个相邻的算子包括能够融合输出的算子与逐元素复用的算子；第一分支为第一深度学习模型的计算图中的任意一个分支；

[0175] 在第一深度学习模型的计算图的至少两个并行分支中，当存在具有相同上游节点的多个算子时，将多个算子组成第一深度学习模型的第二候选算子组合。

[0176] 具体地，可以先根据第一深度学习模型的计算图，确定第一深度学习模型中的各个算子的类型，可以理解的是，相同类型的算子对输入数据执行相同类型的数据处理操作。

[0177] 对于第一深度学习模型的计算图的同一分支内的相邻算子，将满足第一条件、第二条件、第三条件任意一项的至少两个相邻算子确定为第一候选算子组合。具体地：

[0178] 第一条件包括：至少两个相邻的算子均为单射函数算子(injective)。单射函数算子对于某一维度的输入数据，输出相同维度的输出数据，而不改变数据的维度，单射函数算子例如：加法、算数平方根等。例如，算子A1对输入数据加上一个常量，算子A2对输入数据取算数平方根，且算子A1的数据输出是算子A2的数据输入，则算子A1与算子A2可以确定为第一候选算子组合。

[0179] 第二条件包括：至少两个相邻的算子包括约简算子(reduction)以及作为约简算子输入的单射函数算子。约简算子对于第一维度的输入数据，输出第二维度的输出数据，第一维度大于第二维度，可以理解的是，约简算子输入到输出具有降维性质。约简算子例如求和函数算子(sum)、缩放函数算子(scale函数，可用于整体或单方向缩放矩阵元素)等。约简算子与作为约简算子输入的单射函数算子可以确定为第一候选算子组合。例如，算子B1对输入数据取算数平方根，算子B2为求和函数，且算子B1是算子B2的的数据输入，则算子B1与算子B2可以确定为第一算子组合。

[0180] 第三条件包括：至少两个相邻的算子包括能够融合输出的算子(complex-out-

fusable)与逐元素复用的算子(element-wise)。能够融合输出的算子对输入数据执行相应的运算得到多维的可融合的输出数据,例如二维卷积函数算子(conv2d)、批量归一化算子(bn)、线性整流算子(relu)等。逐元素复用的算子是指需要反复对输入数据的全部或部分进行处理的算子。例如,二维卷积函数算子conv2d属于能够融合输出的算子,可以将与逐元素复用算子element-wise的输出与二维卷积函数算子conv2d的输出融合到一起输出,因此可以将二维卷积函数算子conv2d与逐元素复用算子element-wise确定为第一候选算子组合。例如,算子C1是二维卷积函数算子,算子C2为逐元素复用算子,则算子C1与算子C2可以确定为第一算子组合。

[0181] 对于第一深度学习模型的计算图的并行分支中的算子,当至少两个并行分支中存在具有相同上游节点的多个算子,将多个算子组成第一深度学习模型的第二候选算子组合。并行分支中存在具有相同上游节点的多个算子可以是相同的算子,也可以是具有相同输入数据维度的不同算子,需要说明的是相同的算子是指算子的类型相同,并且算子的参数相同,例如两个具有相同卷积核的二维卷积函数算子conv2d。

[0182] 本实施例中根据第一深度学习模型的计算图以及相应的判断规则准确全面地确定了可以用于算子融合的候选算子组合,提升了第一深度学习模型的推理速度。

[0183] 基于上述任一实施例,在一个实施例中,在根据候选算子组合确定对应的目标加速算子之前,方法还包括:

[0184] 根据作为样本的第二深度学习模型的计算图,确定第二深度学习模型的样本算子组合;其中,样本算子组合包括第一样本算子组合和/或第二样本算子组合;第一样本算子组合是第二深度学习模型的计算图中,同一分支内的具有依赖关系的多个样本算子的组合;第二样本算子组合是第二深度学习模型的计算图中,并行分支内具有相同上游节点的多个样本算子的组合;

[0185] 获取样本算子组合中的各个样本算子分别对应的已验证样本算子,并保存在基础算子库中;

[0186] 对样本算子组合中的各个样本算子所对应的已验证样本算子进行融合,得到与样本算子组合所对应的加速算子,并保存在加速算子库中。

[0187] 具体地,在根据候选算子组合确定对应的目标加速算子之前,需要预先根据样本深度学习模型建立加速算子库。根据作为样本的第二深度学习模型的计算图,确定第二深度学习模型的可用于进行算子融合的样本算子组合。样本算子组合可以是第二深度学习模型的计算图中,同一分支内的具有依赖关系的多个样本算子组成的第一样本算子组合,还可以是第二深度学习模型的计算图中,并行分支内具有相同上游节点的多个样本算子组成的第二样本算子组合;

[0188] 在确定样本算子组合后,可以基于异构编程为样本算子组合中的各个样本算子分别创建对应的待验证样本算子,然后进一步对待验证算子进行验证,验证内容可以包括性能和准确度,在待验证样本算子验证合格的情况下,将通过验证的已验证样本算子保存到基础算子库中。进一步地,对样本算子组合中的各个样本算子所对应的已验证样本算子进行融合,得到与样本算子组合所对应的加速算子,经过验证后保存在加速算子库中。通过收集多个样本深度学习模型,执行上述步骤,即可构建基础算子库以及加速算子库,用于提高获取目标加速算子的效率。

[0189] 本实施例中根据作为样本的第二深度学习模型确定了样本算子组合，并生成了相应的已验证样本算子以及加速算子，通过收集多个样本深度学习模型，执行上述步骤，构建了基础算子库以及加速算子库，提高了获取目标加速算子的效率。

[0190] 基于上述任一实施例，在一个实施例中，获取样本算子组合中的各个样本算子分别对应的已验证样本算子，包括：

[0191] 为样本算子组合中的各个样本算子分别创建对应的待验证样本算子；

[0192] 对待验证样本算子进行验证；

[0193] 验证合格后得到样本算子组合中的各个样本算子所对应的已验证样本算子。

[0194] 具体地，对样本算子组合中的各个样本算子分别基于异构编程创建对应的算子核函数，利用编译器进行编译得到样本算子组合中的样本算子对应的待验证样本算子，然后对待验证样本算子进行验证，将验证合格的已验证样本算子保存到基础算子库中。

[0195] 可以理解的是，对于验证未通过的待验证样本算子，需要进行分析调整。具体可以进一步分析确定待验证样本算子的核函数设置线程的数量，在核函数设置线程数量大于待处理数据量的情况下，重新创建待验证样本算子的核函数，重新创建的核函数中设置有线程数判断节点，用于判断核函数的当前线程数，在当前线程数小于待处理数据量的情况下，则对待处理数据进行该算子需要的计算。在当前线程数大于或等于待处理数据量的情况下，则不执行该算子需要的计算，从而避免对数据进行不必要的处理带来误差。

[0196] 本实施例中对样本算子组合中的各个样本算子分别创建对应的待验证样本算子，并进行验证，将验证通过的已验证样本算子保存到基础算子库中，保障了基础算子库中已验证样本算子的可用性。

[0197] 基于上述任一实施例，在一个实施例中，对待验证样本算子进行验证，包括：

[0198] 从标准深度学习框架中调用与待验证样本算子相对应的对比算子；

[0199] 为待验证样本算子与对比算子设置相同的数据输入；

[0200] 获取待验证样本算子对数据输入的第一推理耗时及第一推理结果；

[0201] 获取对比算子对数据输入的第二推理耗时及第二推理结果；

[0202] 在第一推理耗时小于第二推理耗时、且第一推理结果与第二推理结果的最大误差小于第一预设阈值的情况下，确定待验证样本算子验证合格。

[0203] 具体地，调用标准深度学习框架(即现有技术中常用的深度学习框架)如pytorch、tensorflow中与待验证算子对应的对比算子，向待验证算子与对比算子输入相同的数据输入(即待处理数据)，获取待验证样本算子对数据输入的第一推理耗时及第一推理结果，获取对比算子对数据输入的第二推理耗时及第二推理结果。然后，根据待验证算子与对比算子的推理耗时差异验证待验证算子的性能：在第一推理耗时小于第二推理耗时，确定待验证算子的性能验证合格，在第一推理耗时大于或等于第二推理耗时，确定待验证算子的性能验证不合格。根据待验证算子与对比算子的推理结果差异验证待验证算子的准确度：在第一推理结果与第二推理结果的最大误差小于第一预设阈值时，确定待验证样本算子准确度验证合格，在第一推理结果与第二推理结果的最大误差大于或等于第一预设阈值时，确定待验证样本算子准确度验证不合格。第一预设阈值的具体数值根据历史数据，按照精度需求进行设置/调整。例如，可以设置第一预设阈值为1e-4(即10的-4次方)。

[0204] 本实施例中通过调用标准深度学习框架中的对比算子对待验证样本算子进行了

验证，保障了基础算子库中已验证样本算子的可用性。

[0205] 基于上述任一实施例，在一个实施例中，对样本算子组合中的各个样本算子所对应的已验证样本算子进行融合，得到与样本算子组合所对应的加速算子，包括：

[0206] 为样本算子组合中的各个样本算子分别调用对应的已验证样本算子，并对已验证样本算子进行融合，得到与样本算子组合所对应的待验证加速算子；

[0207] 对待验证加速算子进行验证；

[0208] 验证合格后得到与样本算子组合所对应的加速算子，并保存在加速算子库中。

[0209] 具体地，从基础算子库中为样本算子组合中的各个样本算子分别调用对应的已验证样本算子，并进行算子融合，得到样本算子组合的待验证加速算子，然后对待验证加速算子进行验证，通过验证后得到与样本算子组合所对应的加速算子。

[0210] 本实施例中根据基础算子库中的已验证算子创建了待验证加速算子，并进行验证，保障了加速算子库中加速算子的可用性。

[0211] 基于上述任一实施例，在一个实施例中，对待验证加速算子进行性能和准确度的验证，包括：

[0212] 为样本算子组合与待验证加速算子设置相同的数据输入；

[0213] 获取待验证加速算子对数据输入的第三推理耗时及第三推理结果；

[0214] 获取样本算子组合对数据输入的第四推理耗时及第四推理结果；

[0215] 在第三推理耗时小于第四推理耗时、且第三推理结果与第四推理结果的最大误差小于第二预设阈值的情况下，确定待验证加速算子验证合格。

[0216] 具体地，可以向样本算子组合与待验证加速算子输入相同的数据输入(即待处理数据)，获取待验证加速算子对数据输入的第三推理耗时及第三推理结果，获取样本算子组合对数据输入的第四推理耗时及第四推理结果。然后，根据待验证加速算子与样本算子组合的推理耗时验证待验证算子的性能：在第三推理耗时小于第四推理耗时，确定待验证加速算子的性能验证合格，在第三推理耗时大于或等于第四推理耗时，确定待验证加速算子的性能验证不合格。根据待验证加速算子与样本算子组合的推理结果验证待验证加速算子的准确度：在第三推理结果与第四推理结果的最大误差小于第二预设阈值的情况下，确定待验证算子的准确度验证合格，在第三推理结果与第四推理结果的最大误差大于或等于第二预设阈值的情况下，确定待验证算子的准确度验证不合格。

[0217] 对于待验证加速算子验证不通过的情况，可以通过逐步融合算子的方式对待验证加速算子进行分析。可以理解的是，当样本算子组合是根据同一分支内具有依赖关系的多个样本算子确定出的第一候选算子组合时，进行分析时可以直接根据算子数据处理的先后顺序逐步融合样本算子组合中的样本算子，得到逐步融合的加速算子。具体而言，对于同一分支内具有依赖关系的样本算子，融合样本算子的顺序即数据流转的先后顺序。当样本算子组合是根据并行分支内具有相同上游节点的多个样本算子确定出的第二候选算子组合时，进行分析时可以预设顺序逐步融合样本算子组合中的样本算子，得到逐步融合的加速算子。具体而言，对于并行分支内具有相同上游节点的相同样本算子，由于数据是并行处理的，数据处理并无严格的先后处理顺序，可以按照预设顺序逐步融合样本算子，例如，按照多个样本算子的标识序号由小到大的顺序进行逐步融合样本算子。逐步融合样本算子后，依次验证融合后的加速算子，将导致融合后验证不合格的样本算子确定为异常样本算子。

在确定出异常样本算子后，重新生成异常样本算子及相应的融合后的加速算子，具体地，可以基于异构编程重新生成异常样本算子的核函数(即程序语言编写的实现算子的代码)，根据异常样本算子的核函数以及上游算子的核函数得到样本算子组合的核函数，然后进一步对样本算子组合的核函数进行编译即可得到样本算子组合融合后的加速算子。进一步地，重新验证融合异常样本算子后的加速算子的性能和准确度，在验证合格的情况下，继续融合下一样本算子，同理，对于下一异常样本算子重复上述步骤，直至最终融合后得到的待验证加速算子验证合格。

[0218] 本实施例中对待验证加速算子进行了验证，保障了加速算子库中加速算子的可用性。

[0219] 图4是本发明提供的一种深度学习模型的推理方法的一个实施例中的程序流程图，下面结合图4中的程序流程图对本发明的一个优选实施例进行说明：

[0220] 过程一：分析深度学习模型的计算图，确定可以融合的至少一个分立算子组合(分立算子即计算图中可独立完成某种数据处理任务的算子节点，这里的分立算子组合相当于上述的候选算子组合)；

[0221] 过程二：按顺序取出一个分立算子组合，在加速算子库中查找是否存在对应的加速算子。若存在对应的加速算子，则调用加速算子库中已有的加速算子在深度学习模型中替换上述取出的分立算子组合，若加速库中已有的加速算子的性能和准确度满足要求，则保留这一替换的改动，若加速库中已有的加速算子的性能和准确度不满足要求，则分析原因并采取有效措施(比如加速算子的输入和分立算子组合的输入、参数设置不完全一致，需要逐个排查调整为一致，再次验证性能和准确度)；若不存在对应的加速算子，则判断分立算子组合中的各个算子是否都存在于基础算子库，若存在于基础算子库，则使用异构编程实现对应算子组合的融合，若不存在于基础算子库，则先实现分立算子(即上述的待验证算子)，验证性能和准确度后加入基础算子库，之后实现对应算子组合的融合，验证性能和准确度满足要求后加入加速算子库，最后将该融合构建的加速算子替换原深度学习模型中对应的分立算子组合。

[0222] 需要说明的是，上述按顺序取出一个分立算子组合是指按照深度学习模型计算图中算子数据处理的先后顺序取出分立算子组合，即优先融合计算图中的上游算子，另外，对于同时存在于纵向分立算子组合(即，计算图中同一分支内的具有依赖关系的多个算子的组合)，以及横向分立算子组合(即，计算图中并行分支内具有相同上游算子的多个算子的组合)中的算子，则优先对纵向分立算子组合进行融合，纵向融合完成后再次确认是否存在可以横向融合的组合。以图3中的深度学习模型计算图为例进行说明，在确定候选算子组合时，左、中、右三个分支中的算子6(共三个)可以确定为横向分立算子组合，假设此时中间分支的算子6与算子4可以确定为纵向分立算子组合，此时中间分支中的算子6同时存在于横向、纵向分立算子组合中，则优先进行中间分支的算子6与算子4的融合，融合完成后再次确认是否存在可以横向融合的组合。由于计算图中数据处理顺序的原因，可进行纵向融合的组合数量通常大于可进行横向融合的组合数量，优先进行纵向融合可以起到更好的融合后的模型加速效果，并且优先进行纵向融合，后进行横向融合，可以有序融合计算图中的算子，避免混乱。

[0223] 过程三：重复过程二，直到过程一中所有的分立算子组合均得到加速算子的替换。

[0224] 本实施例通过具有通用接口的基础算子库以及加速算子库,有利于提高对相似模型的算子进行推理加速的开发效率。

[0225] 图5是本发明提供的一种深度学习模型的推理方法的一个实施例中预先构建加速算子库的程序流程图,下面结合图5中的程序流程图对本发明的一个优选实施例进行说明:

[0226] 如图5所示,预先收集常用深度学习模型(即神经网络模型),分析可以列出融合的分立算子组合,并设计通用的融合后的算子接口,即加速算子接口,方便以后的调用;

[0227] 算子可以分为四类:

[0228] (1)injective(单射函数算子,比如加法算子,算术平方根算子等);

[0229] (2)reduction(约简函数算子,多到少的映射,输入到输出具有降维性质的,比如sum求和算子);

[0230] (3)complex-out-fusable(能够融合输出的算子比如conv2d二维卷积函数算子、bn批量归一化算子、ln批量归一化算子、relu线性整流算子等);

[0231] (4)opaque(不能被融合的,比如sort排序算子等)。

[0232] “融合算子的通用规则”具体为:

[0233] 对于(1),多个单射算子可以融合成另一个单射算子;

[0234] 对于(2),约简算子可以与输入约简算子的单射算子融合(例如scale和sum);

[0235] 对于(3),能够融合输出的算子,可以与逐元素复用算子进行融合,比如conv2d属于complex-out-fusable,可以把element-wise算子的输出与它的输出融合到一起输出。

[0236] 其中,确定可以融合的算子的方法包括:遍历计算图中的算子节点,对于同一分支存在依赖关系的相邻算子按照上述通用规则来判断算子所属类别及适用的规则,对于满足规则的算子进行融合;若计算图中存在其他并行分支,则判断多个并行分支中,是否存在具有相同上游节点的同样结构的或具有相同输入数据维度的算子,如有,则进行横向融合,将融合前的边连接到融合后算子节点上。直到计算图中没有可以纵向或者横向融合的算子。

[0237] 使用异构编程分别实现分立算子组合中各个算子的核函数,编译得到单个算子,验证性能和准确度后,加入基础算子库;

[0238] 其中,对于自己实现的分立算子验证性能和准确度的方法包括:调用已有的深度学习框架(如pytorch、tensorflow等)中对应的算子,给自己实现的算子和深度学习框架中的算子同样的数据输入,对二者的输出逐元素求差值,并分别统计两个算子的平均计算时间。若元素间的最大的误差小于预设阈值(如1e-4,即10的-4次方),则认为准确度达到要求,否则,准确度没达到要求;若自己实现的分立算子的平均计算时间小于等于已有深度学习框架对应的算子,则认为性能达到要求,否则,性能没达到要求。

[0239] 使用异构编程实现融合后的加速算子,即将分立算子组合中的计算过程进行合并,得到融合后的算子组合核函数,称为加速算子,验证加速算子的性能是否高于融合前的分立算子组合,计算数值的误差(或简称为计算误差)是否小于误差阈值,若性能和准确度均达到要求(即,性能高于加速前的算子,表征性能达到要求;计算误差小于误差阈值,表征准确度达到要求),则将其统一接口并加入加速算子库。

[0240] 其中,对于融合后的算子验证性能和准确度的方法包括:给融合前的分立算子组合及融合后的算子同样的数据输入,对二者的输出逐元素求差值,并分别统计二者的平均计算时间。若元素间的最大误差小于预设阈值(如1e-4,即10的-4次方),则认为准确度达到

要求，否则，准确度没达到要求；若融合后的加速算子的平均计算时间小于融合前的算子组合，则认为性能达到要求，否则，性能没达到要求。

[0241] 其中，融合后的加速算子的“计算误差”是指：在给加速前后的算子同样输入的情况下，对计算结果逐元素求差，取各个元素误差的最大值为加速算子的计算误差。

[0242] 对于算子的性能或准确度验证不合格的情况需要进行分析：

[0243] (1)异构编程实现的分立算子验证不合格：

[0244] 分析可能原因：算子核函数中，设置的线程数量多于实际要处理的数据量，且没有在核函数做对应处理。

[0245] 解决办法：在核函数中做对应的判断处理。在核函数中判断当前设置的线程数量是否小于实际数据数量，若是，则进行计算，否则不进行计算。

[0246] (2)异构编程实现的加速算子验证不合格：

[0247] 分析可能原因：分立算子在融合时，某些计算步骤融合为一个后，计算过程与正确计算过程有出入，导致后续计算使用的数据是有误的。

[0248] 解决办法：在融合后的算子核函数中逐个验证分立算子的计算结果和性能，在前一个分立算子性能和误差达到要求的情况下，再将下一个算子的计算步骤逐个融合进去，排查定位问题原因，调整该融合步骤的实现。直至融合后的算子组合准确度及性能达到要求。

[0249] 本实施例通过重写算子，构建基础算子库，对特定算子组合实现算子融合，可以对模型中的算子进行选择性优化加速；构建具有通用接口的算子库，有利于提高对相似模型的算子进行推理加速的开发效率。

[0250] 图6是本发明提供的一种深度学习模型的推理方法中算子融合前的计算图示例之一，图7是本发明提供的一种深度学习模型的推理方法中算子融合后的计算图示例之二。下面结合图6以及图7对本发明的一个优选实施例中同一分支内的算子融合过程进行说明：

[0251] 图6中每一个圆角矩形框代表了一个通常意义的计算过程，不同框架的底层有不同的实现，具体到cuda kernel级别而言，可能是多个kernel(算子运算节点)的有序组合得到一个矩形框中的计算过程。图3中的每一个矩形框则代表cuda(通用并行计算架构)装置中的一个计算kernel。

[0252] 观察图6，可以发现add bias和activation两个算子的计算过程是相邻的，且均属于element-wise计算，即逐元素的计算。这两个运算可以放在一个kernel中计算，即融合在一起成为一个计算kernel。类似地，add bias、add residual、layer normalization三个算子计算过程也是依次相邻的，并且layer normalization(即ln，批量归一化算子)用到的数据是前一步add residual产生的，因此其可以和前两者进行融合。

[0253] 图7直观地展示了融合后的算子组合的计算过程，可以明显的看到，融合算子后的计算过程相比加速前有所简化，这种简化具体到计算底层，意味着减少了数据的读写，也就意味着减少了推理所需时间，从而实现推理加速。

[0254] 本实施例中通过算子融合减少了推理所需时间，实现了推理加速。

[0255] 图8是本发明提供的一种深度学习模型的推理方法中算子融合前的计算图示例之三，图9是本发明提供的一种深度学习模型的推理方法中算子融合后的计算图示例之四，图10是本发明提供的一种深度学习模型的推理方法中算子融合后的计算图示例之五。下面结

合图8-10对本发明的一个优选实施例中算子融合过程进行说明：

[0256] 可以看到，图8是融合前的原始深度学习模型的计算图，图中concat是输出节点，input是输入节点，next input是下一输入节点，relu、bias、conv、max pool是计算图中的算子节点。图8到图9是对竖直方向上的四条分支，分别在分支内作了relu+bias+conv算子融合。图9到图10，进一步在图9的基础上对水平方向上并行分支中的1x1 CBR算子进行了融合，将所有1x1的CBR融合成一个大的CBR。

[0257] 本实施例中通过算子融合减少了推理所需时间，实现了推理加速。

[0258] 本发明收集现有常用的深度学习模型，以现有常用的深度学习模型为对象预先设计加速后的通用加速算子，统一加速算子的接口，构建形成具有通用接口的加速算子库，有利于提高对相似模型的算子进行推理加速的开发效率。对于待处理的深度学习模型，分析待处理的深度学习模型的计算图，如果发现模型中存在可以融合的算子，并且加速算子库中已经有相关融合后的加速算子，则可以使用加速算子库中的融合后的加速算子替换原有候选算子组合，从而实现对深度学习模型的推理加速。

[0259] 基于上述构思，本发明可以应用于多种涉及深度学习的应用场景中，提升深度学习模型的推理速度。例如，应用于图像识别应用场景中，根据本发明提供的方法对图像识别深度学习模型进行算子融合，可以提升图像识别深度学习模型的图像识别速度。又例如，应用于机器翻译领域，根据本发明提供的方法对机器翻译深度学习模型进行算子融合，可以提升机器翻译深度学习模型的机器翻译。速度。又例如，应用于情感识别领域，根据本发明提供的方法对机器翻译深度学习模型进行算子融合，可以提升情感深度学习模型的情感识别速度。

[0260] 下面对本发明提供的一种机器翻译方法进行描述，下文描述的机器翻译方法与上文描述的深度学习模型的推理方法可相互对应参照。

[0261] 图11是本发明提供的一种机器翻译方法的流程示意图，如图1所示，该方法包括：

[0262] T110、将待翻译文本输入第一机器翻译模型，得到翻译后的文本；

[0263] 其中，第一机器翻译模型是利用目标加速算子替换初始机器翻译模型中相应的候选算子组合后得到的；第一机器翻译模型的获取过程包括：

[0264] 根据初始机器翻译模型的计算图，确定初始机器翻译模型的候选算子组合；其中，候选算子组合包括第一候选算子组合和/或第二候选算子组合；第一候选算子组合是第一机器翻译模型的计算图中，同一分支内的具有依赖关系的多个算子的组合；第二候选算子组合是第一机器翻译模型的计算图中，并行分支内具有相同上游节点的多个算子的组合；

[0265] 根据候选算子组合确定对应的目标加速算子；其中，目标加速算子包括第一加速算子和/或第二加速算子；第一加速算子是作为样本的第二机器翻译模型的计算图中，同一分支内的具有依赖关系的多个算子融合后得到的；第二加速算子是第二机器翻译模型的计算图中；并行分支内具有相同上游节点的多个算子融合后得到的；

[0266] 将目标加速算子替换初始机器翻译模型中的候选算子组合，得到第一机器翻译模型。

[0267] 具体地，首先，根据初始机器翻译模型的计算图，确定出初始机器翻译模型的可以进行算子融合的候选算子组合。机器翻译模型是用于实现机器翻译的深度学习模型。计算图用于将计算过程图形化，机器翻译模型将输入的待翻译数据经过一系列数据处理得到翻

译结果，机器翻译模型中的对数据执行的各种处理操作即可抽象为算子，算子对应于计算图中的算子节点，可以理解的是计算图中的算子节点可以是独立整体的单个算子，也可以是由更细粒度的算子的有序组合；机器翻译模型中的各步骤之间的数据流转可抽象为算子之间的依赖关系(即一个算子的输出或部分输出是另一个算子的输入或部分输入)，算子之间的依赖关系对应于计算图中的算子节点之间的连接关系。候选算子组合可以是初始机器翻译模型的计算图中同一分支内的具有依赖关系的多个算子组成的第一候选算子组合。举例说明如下：请参照图2，如图2所示，算子1、算子2、算子3处于初始机器翻译模型的计算图的同一分支，且依次具有依赖关系，此时第一候选算子组合可以是{算子1，算子2}，可以是{算子2，算子3}，还可以是{算子1，算子2，算子3}。同一分支内具有依赖关系的算子之间的输入数据输出数据存在关联，将其确定为候选算子组合用于进行后续算子融合，可以有效减少第一机器翻译模型对于数据的存取次数，提升模型推理效率。

[0268] 候选算子组合还可以是初始机器翻译模型的计算图中并行分支内具有相同上游节点的多个算子组成的第二候选算子组合，其中“多个”是指“至少两个”。举例说明如下：请参照图3，如图3所示，计算图中包括左、中、右三个并行分支，以及算子1-算子6共六种算子，此时，第二候选算子组合可以是左、中两个并行分支中的两个算子4组成的算子组合，还可以是左、中、右三个并行分支中的三个算子6组成的算子组合。并行分支内具有相同上游节点的多个算子意味着该多个算子的数据输入节点相同，将其确定为候选算子组合用于进行后续算子融合，可以有效减少初始机器翻译模型对于数据的存取次数，提升模型推理效率。

[0269] 进一步地，并行分支内具有相同上游节点的多个算子可以是相同的算子，相同的算子执行的数据处理内容相同，因此，并行分支内具有相同上游节点的多个相同算子可以组成第二候选算子组合，例如图3中所示的左、中、右三个并行分支中的三个算子6可以组成第二候选算子组合。

[0270] 并行分支内具有相同上游节点的多个算子还可以是输入数据维度相同的不同算子，举例来说，第一分支的算子b_1与第二分支的算子b_2对相同的上游节点a输入的二维数据进行处理，b_1为二维卷积算子，b_2为四维卷积算子，则b_2可以拆分出一个二维卷积算子，可以将b_2拆出的二维卷积算子部分与b_1进行融合，可见，b_1，b_2也可以确定为第二候选算子组合。相同数据输入数据的维度相同才能保证融合的线程相同，进而可以进行算子融合，减少融合后算子的种类。

[0271] 然后，根据候选算子组合确定对应的目标加速算子。目标加速算子可以是根据作为样本的第二机器翻译模型的计算图中同一分支内的具有依赖关系的多个算子融合后得到的第一加速算子，还可以是第二机器翻译模型的计算图中并行分支内具有相同上游节点的多个算子融合后得到的第二加速算子。依赖关系是指一个算子的输出(或者输出的一部分)是另一个算子的输入(或者输入的一部分)，此时这两个算子存在依赖关系。算子融合的目的是减少整个计算图中算子节点的数量，对应到推理设备来说，从一个算子节点到另一个算子节点之间就意味着数据的搬运，减少算子节点相应地可以减少数据的存取调用，进而提升推理设备的工作效率。算子融合的实现方式可以是：根据待融合的算子组合中的各个算子，基于异构编程(如CPU+GPU)生成各个算子的核函数(即程序语言编写的实现算子的代码)，并根据各个算子的核函数生成算子组合的核函数，然后进一步利用编译器(如nvcc编译器)对算子组合的核函数进行编译即可得到算子组合融合后的加速算子。需要说明的

是,机器翻译模型的训练/推理过程通常涉及大量的并行运算,基于这一特性,利用异构编程实现的机器翻译模型可以部署于异构计算系统,从而大幅提升机器翻译模型的训练/推理效率。异构计算系统通常由通用处理器和许多特定于域的处理器组成:通用处理器作为控制设备(称为主机),用于复杂的控制和调度;特定于域的处理器作为子设备(称为MLU),用于大规模并行计算和特定于域的计算任务。主机和MLU合作完成计算任务。对于异构计算系统,原始的同构并行编程模型不再适用,因此需要用到异构编程。算子融合的实现方式还可以是:基于同构编程生成算子组合的核函数,并经过编译后得到相应的加速算子,可以理解的是,基于同构编程实现的机器翻译模型需要部署于同构计算系统,但是相比于利用异构编程实现并部署于异构计算系统的机器翻译模型,性能有一定程度降低。

[0272] 将目标加速算子替换初始机器翻译模型中的候选算子组合,得到第一机器翻译模型,用于进行机器翻译

[0273] 本实施例中根据初始机器翻译模型的计算图中同一分支内的具有依赖关系的算子或并行分支内具有相同上游节点的算子确定候选算子组合,在算子层面精确、灵活地确定了可融合的算子;根据候选算子组合确定了初始机器翻译模型的加速算子;通过将初始机器翻译模型中的候选算子组合替换为加速算子,得到第一机器翻译模型,提升了第一机器翻译模型的推理效率,实现灵活高效地对模型进行推理加速,提升了用户体验,降低了推理设备成本。

[0274] 基于上述实施例,在一个实施例中,根据候选算子组合确定对应的目标加速算子,包括:根据候选算子组合在预先建立的加速算子库中查找对应的目标加速算子;其中,加速算子库包括算子组合与加速算子之间的映射关系;或,将所选算子组合中的算子进行融合,得到对应的目标加速算子。

[0275] 具体地,可以基于预先建立的加速算子库中查找对应的目标加速算子,加速算子库中包括,作为样本的第二深度学习模型的计算图中同一分支内的具有依赖关系的多个算子融合后得到的第一加速算子,以及第二深度学习模型的计算图中,并行分支内具有相同上游节点的多个算子融合后得到的第二加速算子。加速算子库中还包括算子组合与加速算子之间的映射关系,根据第一深度学习模型的候选算子组合,结合映射关系即可确定候选算子组合对应的目标加速算子。通过预先收集作为样本的第二深度学习模型,丰富加速算子库中的加速算子,可以有效提升第一深度学习模型的确定加速算子的效率。

[0276] 还可以基于候选算子组合中的算子直接进行算子融合,得到对应的目标加速算子。

[0277] 本实施例中根据候选算子组合,基于预先建立的算子库或直接进行算子融合,全面准确地确定了目标加速算子,进一步提高了深度学习模型的推理效率。

[0278] 基于上述任一实施例,在一个实施例中,将目标加速算子替换初始机器翻译模型中的候选算子组合,得到第一机器翻译模型,包括:

[0279] 将初始机器翻译模型中的候选算子组合替换为目标加速算子;

[0280] 根据替换后的初始机器翻译模型进行推理测试,根据推理测试的结果对目标加速算子进行验证;

[0281] 在目标加速算子验证合格的情况下,将替换后的初始机器翻译模型确定为第一机器翻译模型;

[0282] 在目标加速算子验证不合格的情况下，对目标加速算子进行分析，根据分析结果调整目标加速算子，然后将第一机器翻译模型中的候选算子组合替换为调整后的目标加速算子，并重新执行根据替换后的第一机器翻译模型进行推理测试的步骤。

[0283] 具体地，将替换后的初始机器翻译模型用于进行推理之前，还需要对替换后的初始机器翻译模型进行验证，验证目标加速算子的加速效果，在目标加速算子验证合格，即目标加速算子的性能满足性能预设条件且目标加速算子的准确度满足准确度预设条件的情况下才可以用于替换进行推理翻译。推理测试的具体内容可以是：在相同推理设备条件下，对替换前后的初始机器翻译模型给予相同的数据输入，记录推理结果以及推理耗时。目标加速算子的性能的预设条件例如：替换后的推理耗时的平均值小于与替换前的推理耗时的平均值；目标加速算子的准确度预设条件例如：替换后的输出结果与替换前的输出结果之间的最大误差小于预设阈值。在目标加速算子的性能和准确度满足预设条件的情况下，即可将替换后的初始机器翻译模型确定为第一机器翻译模型，用于进行推理翻译。

[0284] 在目标加速算子验证不合格，即目标加速算子的性能不满足性能预设条件或准确度不满足准确度预设条件，则说明将候选算子组合替换为目标加速算子会导致最终得到的第一机器翻译模型的推理速度得不到提升或者推理的准确率降低，因此，对于目标加速算子验证不合格的情况，需要对目标加速算子进行分析，根据分析结果重新调整目标加速算子，然后将初始机器翻译模型中的候选算子组合替换为调整后的目标加速算子，并重新执行根据替换后的初始机器翻译模型进行推理测试的步骤，直至得到验证合格的目标加速算子。

[0285] 本实施例中，在将替换后的初始机器翻译模型用于进行机器翻译之前，对替换后的初始机器翻译模型进行推理测试，验证目标加速算子的加速效果，保障了最终得到的第一机器翻译模型的翻译准确率以及翻译加速效果，在目标加速算子验证不合格的情况下，对目标加速算子进行分析，根据分析结果调整加速算子，保障了最终得到的第一深度学习模型的推理准确率，并提升了推理速度。

[0286] 基于上述任一实施例，在一个实施例中，对目标加速算子进行分析，根据分析结果调整目标加速算子，包括：

[0287] 当目标加速算子是第一加速算子时根据算子数据处理的先后顺序，逐步融合第一候选算子组合中的算子，或，当目标加速算子是第二加速算子时，根据预设顺序逐步融合第二候选算子组合中的算子；

[0288] 依次验证融合后的加速算子，确定出导致融合后的加速算子验证不合格的异常算子；

[0289] 重新构建异常算子，并验证融合重新构建的异常算子后的加速算子，在验证合格的情况下，继续融合下一算子，直至最终融合后的目标加速算子验证合格。

[0290] 具体地，可以通过逐步融合算子的方式对目标加速算子进行分析，可以理解的是，当目标加速算子是上述第一加速算子时，此时目标加速算子对应的候选算子组合是根据同一分支内具有依赖关系的多个算子确定出的第一候选算子组合，进行分析时可以直接根据算子数据处理的先后顺序逐步融合第一候选算子组合中的算子。具体而言，对于同一分支内具有依赖关系的算子，融合算子的顺序即数据流转的先后顺序。当目标加速算子是第二加速算子时，此时目标加速算子对应的候选算子组合是根据并行分支内具有相同上游节点

的多个算子确定出的第二候选算子组合，由于数据是并行处理的，数据处理并无严格的先后处理顺序，进行分析时可以按照预设的顺序进行逐步融合算子，预设的顺序例如多个算子的标识序号由小到大的顺序。

[0291] 逐步融合算子后，依次验证融合后的加速算子，验证内容可以包括融合算子的性能和准确度，将导致融合后性能或准确度不满足预设条件的算子确定为异常算子。在确定出异常算子后，重新生成异常算子及相应的融合后的加速算子，具体地，可以基于异构编程重新生成异常算子的核函数(即程序语言编写的实现算子的代码)，对异常算子逐步排查计算结果，对有问题的计算步骤改写为正确的计算步骤，根据重新生成的异常算子的核函数以及上游算子的核函数得到算子组合的核函数，然后进一步对算子组合的核函数进行编译即可得到算子组合再次融合后的加速算子。进一步地，重新验证融合异常算子后的加速算子，在验证合格的情况下，继续融合下一算子，同理，对于下一异常算子重复上述步骤，直至最终融合后得到的目标加速算子的性能和准确度满足预设条件。

[0292] 本实施例中，通过逐步融合算子并进行验证确定出了导致目标加速算子验证不合格的异常算子，对异常算子逐步排查计算结果，对有问题的计算步骤改写为正确的计算步骤，通过重新生成异常算子以及融合异常算子后的加速算子并进行验证，对目标加速算子重新生成并验证性能及准确性，进一步保障了替换后得到的第一机器翻译模型的翻译翻译结果与替换前的推理结果的一致性，并提升了翻译速度。

[0293] 基于上述任一实施例，在一个实施例中，在根据候选算子组合在预先建立的加速算子库中查找对应的目标加速算子之后，在将目标加速算子替换初始机器翻译模型中的候选算子组合之前，方法还包括：

[0294] 在未为候选算子组合查找到对应目标加速算子的情况下，针对候选算子组合中的每个算子执行以下处理：在预先建立的基础算子库中查找对应算子的已验证算子；

[0295] 将与候选算子组合中的多个算子一一对应的多个已验证算子进行融合，得到与候选算子组合所对应的目标加速算子。

[0296] 具体地，本实施例中还预先建立了基础算子库，基础算子库中包括多个可用于进行算子融合的预先经过验证的已验证算子。在无法利用加速算子库直接获取候选算子组合对应的目标加速算子的情况下，可以利用预先建立的基础算子库生成目标加速算子。若候选算子组合中的每个算子经查找确定在基础算子库都存在所对应的已验证算子，此时可以直接利用基础算子库中每个算子对应的已验证算子进行融合得到与候选算子组合所对应的目标加速算子，提高获取目标加速算子的效率。

[0297] 本实施例中，通过基础算子库，提高了获取目标加速算子的效率。

[0298] 基于上述任一实施例，在一个实施例中，

[0299] 当在预先建立的基础算子库中未查找到对应算子的已验证算子时，方法还包括：

[0300] 创建对应算子的待验证算子，并对待验证算子进行验证，将验证合格的算子确定为算子对应的已验证算子。

[0301] 具体地，对于候选算子组合中的算子，若在基础算子库中不存在对应于该算子的已验证算子，则需要重新创建待验证算子。具体地，可以基于异构编程生成该算子的核函数(即程序语言编写的实现算子的代码)，经过编译器编译后得到重新创建的待验证算子，并对重新创建的待验证算子进行验证，验证内容可以包括性能和准确度。验证过程可以通过

如下方式实现:为待验证样本算子与对比算子设置相同的数据输入,并记录待验证样本算子与对比算子的推理结果、推理耗时。根据推理结果验证待验证样本算子准确度是否满足准确度预设条件,根据推理耗时验证待验证样本算子的性能是否满足性能预设条件。对比算子可以是已有的深度学习框架如pytorch、tensorflow中与待验证算子对应的对比算子,性能合格的预设条件可以是待验证样本算子的推理耗时小于对比算子的推理耗时,准确度合格的预设条件可以是待验证样本算子的推理结果与对比算子的推理结果之间的最大误差小于预设阈值(例如1e-4,即10的-4次方)。在待验证算子满足预设条件的情况下,将已验证算子加入基础算子库中。重复上述步骤,直至在基础算子库中可以查找到候选算子组合中的各个算子所对应的已验证算子。

[0302] 本实施例中,在基础算子库中不存在与候选算子组合中的算子对应的已验证算子的情况下,分别创建并验证待验证算子,便于顺利地获取目标加速算子。

[0303] 基于上述任一实施例,在一个实施例中,根据初始机器翻译模型的计算图,确定第一机器翻译模型的候选算子组合,包括:

[0304] 根据初始器翻译模型的计算图,确定初始机器翻译模型中的各个算子的类型;

[0305] 在初始机器翻译模型的计算图的第一分支中,当至少两个相邻的算子满足第一条件、第二条件与第三条件中的任意一项时,则将至少两个相邻的算子组成初始机器翻译模型的第一候选算子组合;其中,第一条件包括:至少两个相邻的算子均为单射函数算子;第二条件包括:至少两个相邻的算子包括约简算子以及作为约简算子输入的单射函数算子;第三条件包括:至少两个相邻的算子包括能够融合输出的算子与逐元素复用的算子;第一分支为初始机器翻译模型的计算图中的任意一个分支;

[0306] 在初始机器翻译模型的计算图的至少两个并行分支中,当存在具有相同上游节点的多个算子,将多个算子组成初始机器翻译模型的第二候选算子组合。

[0307] 具体地,可以先根据初始机器翻译模型的计算图,确定初始机器翻译模型中的各个算子的类型,相同类型的算子对输入数据执行相同类型的数据处理操作。

[0308] 对于初始机器翻译模型的计算图的同一分支内的相邻算子,将满足第一条件、第二条件、第三条件任意一项的至少两个相邻算子确定为第一候选算子组合。具体地:

[0309] 第一条件包括:至少两个相邻的算子均为单射函数算子(injective)。单射函数算子对于某一维度的输入数据,输出相同维度的输出数据,而不改变数据的维度,单射函数算子例如:加法、算数平方根等。例如,算子A1对输入数据加上一个常量,算子A2对输入数据取算数平方根,且算子A1的数据输出是算子A2的数据输入,则算子A1与算子A2可以确定为第一候选算子组合。

[0310] 第二条件包括:至少两个相邻的算子包括约简算子(reduction)以及作为约简算子输入的单射函数算子。约简算子对于第一维度的输入数据,输出第二维度的输出数据,第一维度大于第二维度,可以理解的是,约简算子输入到输出具有降维性质。约简算子例如求和函数算子(sum)、缩放函数算子(scale函数,可用于整体或单方向缩放矩阵规模)等。约简算子与作为约简算子输入的单射函数算子可以确定为第一候选算子组合。例如,算子B1对输入数据取算数平方根,算子B2为求和函数,且算子B1是算子B2的的数据输入,则算子B1与算子B2可以确定为第一算子组合。

[0311] 第三条件包括:至少两个相邻的算子包括能够融合输出的算子(complex-out-

fusable)与逐元素复用的算子(element-wise)。能够融合输出的算子对输入数据执行相应的运算得到多维的可融合的输出数据,例如二维卷积函数算子(conv2d)、批量归一化算子(bn)、线性整流算子(relu)等。逐元素复用的算子是指需要反复对输入数据的全部或部分进行处理的算子。例如,二维卷积函数算子conv2d属于能够融合输出的算子,可以将与逐元素复用算子element-wise的输出与二维卷积函数算子conv2d的输出融合到一起输出,因此可以将二维卷积函数算子conv2d与逐元素复用算子element-wise确定为第一候选算子组合。例如,算子C1是二维卷积函数算子,算子C2为逐元素复用算子,则算子C1与算子C2可以确定为第一算子组合。

[0312] 对于初始机器翻译模型的计算图的并行分支中的算子,当至少两个并行分支中存在具有相同上游节点的多个算子,将多个算子组成初始机器翻译模型的第二候选算子组合。并行分支中存在具有相同上游节点的多个算子可以是相同的算子,也可以是具有相同输入数据维度的不同算子,需要说明的是相同的算子是指算子的类型相同,并且算子的参数相同,例如两个具有相同卷积核的二维卷积函数conv2d。

[0313] 本实施例中根据初始机器翻译模型的计算图以及相应的判断规则准确全面地确定了可以用于算子融合的候选算子组合,提升了得到的第一机器翻译模型的翻译速度。

[0314] 基于上述任一实施例,在一个实施例中,在根据候选算子组合确定对应的加速算子之前,方法还包括:

[0315] 根据作为样本的第二机器翻译模型的计算图,确定第二机器翻译模型的样本算子组合;其中,样本算子组合包括第一样本算子组合和/或第二样本算子组合;第一样本算子组合是第二机器翻译模型的计算图中,同一分支内的具有依赖关系的多个样本算子的组合;第二样本算子组合是第二机器翻译模型的计算图中,并行分支内具有相同上游节点的多个样本算子的组合;

[0316] 获取样本算子组合中的各个样本算子分别对应的已验证样本算子,并保存在基础算子库中;

[0317] 对样本算子组合中的各个样本算子所对应的已验证样本算子进行融合,得到与样本算子组合所对应的加速算子,并保存在加速算子库中。

[0318] 具体地,在根据候选算子组合在预先建立的加速算子库中查找对应的目标加速算子之前,需要预先根据样本机器翻译深度学习模型建立加速算子库。根据作为样本的第二机器翻译模型的计算图,确定第二机器翻译模型的可用于进行算子融合的样本算子组合。样本算子组合可以是第二机器翻译模型的计算图中,同一分支内的具有依赖关系的多个样本算子组成的第一样本算子组合,还可以是第二机器翻译模型的计算图中,并行分支内具有相同上游节点的多个样本算子组成的第二样本算子组合;

[0319] 在确定样本算子组合后,可以基于异构编程为样本算子组合中的各个样本算子分别创建对应的待验证样本算子,然后进一步对待验证样本算子进行验证,验证内容可以包括性能和准确度,在待验证样本算子验证合格的情况下,将通过验证的已验证样本算子保存到基础算子库中。进一步地,对样本算子组合中的各个样本算子所对应的已验证样本算子进行融合,得到与样本算子组合所对应的加速算子,经过验证后保存在加速算子库中。通过收集多个样本机器翻译模型,执行上述步骤,即可构建基础算子库以及加速算子库,用于提高获取初始机器翻译模型的目标加速算子的效率。

[0320] 本实施例中根据作为样本的第二机器翻译模型确定了样本算子组合，并生成了相应的已验证样本算子以及加速算子，通过收集多个样本机器翻译模型，执行上述步骤，构建了基础算子库以及加速算子库，提高了获取初始机器翻译模型的目标加速算子的效率。

[0321] 基于上述任一实施例，在一个实施例中，获取样本算子组合中的各个样本算子分别对应的已验证样本算子，包括：

[0322] 为样本算子组合中的各个样本算子分别创建对应的待验证样本算子；

[0323] 对待验证样本算子进行验证；

[0324] 验证合格后得到样本算子组合中的各个样本算子所对应的已验证样本算子。

[0325] 具体地，对样本算子组合中的各个样本算子分别基于异构编程创建对应的算子核函数，利用编译器进行编译得到样本算子组合中的样本算子对应的待验证样本算子，然后对待验证样本算子进行验证，将验证合格的已验证样本算子保存到基础算子库中。

[0326] 可以理解的是，对于验证未通过的待验证样本算子，需要进行分析调整。具体可以进一步分析确定待验证样本算子的核函数设置线程的数量，在核函数设置线程数量大于待处理数据量的情况下，重新创建待验证样本算子的核函数，重新创建的核函数中设置有线程数判断节点，用于判断核函数的当前线程数，在当前线程数小于待处理数据量的情况下，则对待处理数据量进行该算子需要的计算。在当前线程数大于或等于待处理数据量的情况下，则不执行该算子需要的计算，从而避免对数据进行不必要的处理带来误差。

[0327] 本实施例中对样本算子组合中的各个样本算子分别创建对应的待验证样本算子，并进行验证，将验证通过的已验证样本算子保存到基础算子库中，保障了基础算子库中已验证样本算子的可用性。

[0328] 基于上述任一实施例，在一个实施例中，对待验证样本算子进行验证，包括：

[0329] 从标准深度学习框架中调用与待验证样本算子相对应的对比算子；

[0330] 为待验证样本算子与对比算子设置相同的操作，验证待验证样本算子的性能和准确度。

[0331] 具体地，调用已有的深度学习框架(即现有技术中常用的深度学习框架)如pytorch、tensorflow中与待验证算子对应的对比算子，向待验证算子与对比算子输入相同的数据输入(即待处理数据)，获取待验证样本算子对数据输入的第一推理耗时及第一推理结果，获取对比算子对数据输入的第二推理耗时及第二推理结果。然后，根据待验证算子与对比算子的推理耗时差异验证待验证算子的性能：在第一推理耗时小于第二推理耗时，确定待验证算子的性能验证合格，在第一推理耗时大于或等于第二推理耗时，确定待验证算子的性能验证不合格。根据待验证算子与对比算子的推理结果差异验证待验证算子的准确度：在第一推理结果与第二推理结果的最大误差小于第一预设阈值时，确定待验证样本算子准确度验证合格，在第一推理结果与第二推理结果的最大误差大于或等于第一预设阈值时，确定待验证样本算子准确度验证不合格。第一预设阈值的具体数值根据历史数据，按照精度需求进行设置/调整。例如，可以设置第一预设阈值为1e-4(即10的-4次方)。本实施例中通过调用标准深度学习框架中的对比算子对待验证样本算子进行了验证验证，保障了基础算子库中已验证样本算子的可用性。

[0332] 基于上述任一实施例，在一个实施例中，对样本算子组合中的各个样本算子所对应的已验证样本算子进行融合，得到与样本算子组合所对应的加速算子，包括：

[0333] 为样本算子组合中的各个样本算子分别调用对应的已验证样本算子，并对已验证样本算子进行融合，得到与样本算子组合所对应的待验证加速算子；

[0334] 对待验证加速算子进行验证；

[0335] 验证合格后得到与样本算子组合所对应的加速算子，并保存在加速算子库中。

[0336] 具体地，从基础算子库中为样本算子组合中的各个样本算子分别调用对应的已验证样本算子，并进行算子融合，得到样本算子组合的待验证加速算子，然后对待验证加速算子进行验证，通过验证后得到与样本算子组合所对应的加速算子。

[0337] 本实施例中根据基础算子库中的已验证算子创建了待验证加速算子，并进行验证，保障了加速算子库中加速算子的可用性。

[0338] 基于上述任一实施例，在一个实施例中，对待验证加速算子进行性能和准确度的验证，包括：

[0339] 为样本算子组合与待验证加速算子设置相同的数据输入；

[0340] 获取待验证加速算子对数据输入的第三推理耗时及第三推理结果；

[0341] 获取样本算子组合对数据输入的第四推理耗时及第四推理结果；

[0342] 在第三推理耗时小于第四推理耗时、且第三推理结果与第四推理结果的最大误差小于第二预设阈值的情况下，确定待验证加速算子验证合格。

[0343] 具体地，可以向样本算子组合与待验证加速算子输入相同的数据输入(即待处理数据)，获取待验证加速算子对数据输入的第三推理耗时及第三推理结果，获取样本算子组合对数据输入的第四推理耗时及第四推理结果。然后，根据待验证加速算子与样本算子组合的推理耗时验证待验证算子的性能：在第三推理耗时小于第四推理耗时，确定待验证加速算子的性能验证合格，在第三推理耗时大于或等于第四推理耗时，确定待验证加速算子的性能验证不合格。根据待验证加速算子与样本算子组合的推理结果验证待验证加速算子的准确度：在第三推理结果与第四推理结果的最大误差小于第二预设阈值的情况下，确定待验证算子的准确度验证合格，在第三推理结果与第四推理结果的最大误差大于或等于第二预设阈值的情况下，确定待验证算子的准确度验证不合格。

[0344] 对于待验证加速算子验证不通过的情况，可以通过逐步融合算子的方式对待验证加速算子进行分析。可以理解的是，当样本算子组合是根据同一分支内具有依赖关系的多个样本算子确定出的第一候选算子组合时，进行分析时可以直接根据算子数据处理的先后顺序逐步融合样本算子组合中的样本算子，得到逐步融合的加速算子。具体而言，对于同一分支内具有依赖关系的样本算子，融合样本算子的顺序即数据流转的先后顺序。当样本算子组合是根据并行分支内具有相同上游节点的多个样本算子确定出的第二候选算子组合时，进行分析时可以预设顺序逐步融合样本算子组合中的样本算子，得到逐步融合的加速算子。具体而言，对于并行分支内具有相同上游节点的相同样本算子，由于数据是并行处理的，数据处理并无严格的先后处理顺序，可以按照预设顺序逐步融合样本算子，例如，按照多个样本算子的标识序号由小到大的顺序进行逐步融合样本算子。逐步融合样本算子后，依次验证融合后的加速算子，将导致融合后验证不合格的样本算子确定为异常样本算子。在确定出异常样本算子后，重新生成异常样本算子及相应的融合后的加速算子，具体地，可以基于异构编程重新生成异常样本算子的核函数(即程序语言编写的实现算子的代码)，根据异常样本算子的核函数以及上游算子的核函数得到样本算子组合的核函数，然后进一步

对样本算子组合的核函数进行编译即可得到样本算子组合融合后的加速算子。进一步地，重新验证融合异常样本算子后的加速算子的性能和准确度，在验证合格的情况下，继续融合下一样本算子，同理，对于下一异常样本算子重复上述步骤，直至最终融合后得到的待验证加速算子验证合格。

[0345] 本实施例中对待验证加速算子进行了验证，保障了加速算子库中加速算子的可用性。

[0346] 下面对本发明提供的深度学习模型的推理装置进行描述，下文描述的深度学习模型的推理装置与上文描述的深度学习模型的推理方法可相互对应参照。

[0347] 图12是本发明提供的一种深度学习模型的推理装置的结构示意图，如图12所示，该装置包括：

[0348] 候选算子确定模块121，用于根据待处理的第一深度学习模型的计算图，确定第一深度学习模型的候选算子组合；其中，候选算子组合包括第一候选算子组合和/或第二候选算子组合；第一候选算子组合是第一深度学习模型的计算图中，同一分支内的具有依赖关系的多个算子的组合；第二候选算子组合是第一深度学习模型的计算图中，并行分支内具有相同上游节点的多个算子的组合；

[0349] 加速算子获取模块122，用于根据候选算子组合确定对应的目标加速算子；其中，目标加速算子包括第一加速算子和/或第二加速算子；第一加速算子是作为样本的第二深度学习模型的计算图中，同一分支内的具有依赖关系的多个算子融合后得到的；第二加速算子是第二深度学习模型的计算图中，并行分支内具有相同上游节点的多个算子融合后得到的；

[0350] 加速算子替换模块123，用于将目标加速算子替换第一深度学习模型中的候选算子组合，根据替换后的第一深度学习模型进行推理。

[0351] 本实施例中根据待处理的第一深度学习模型的计算图中同一分支内的具有依赖关系的算子或并行分支内具有相同上游节点的算子确定候选算子组合，在算子层面精确、灵活地确定了可融合的算子；根据候选算子组合确定了第一深度学习模型的加速算子；通过将候选算子组合替换为加速算子，提升了第一深度学习模型的推理效率，实现灵活高效地对模型进行推理加速，提升了用户体验，降低了推理设备成本。

[0352] 基于上述任一实施例，在一个实施例中，所述加速算子获取模块122包括：

[0353] 第一获取单元，用于根据所述候选算子组合在预先建立的加速算子库中查找对应的目标加速算子；其中，所述加速算子库包括算子组合与加速算子之间的映射关系；

[0354] 第二获取单元，用于将所述候选算子组合中的算子进行融合，得到对应的目标加速算子。

[0355] 基于上述任一实施例，在一个实施例中，加速算子替换模块123，包括：

[0356] 第一替换单元，用于将第一深度学习模型中的候选算子组合替换为目标加速算子；

[0357] 第二替换单元，用于根据替换后的第一深度学习模型进行推理测试，根据推理测试的结果对目标加速算子进行验证；

[0358] 第三替换单元，用于在目标加速算子验证合格的情况下，根据替换后的第一深度学习模型进行推理；

[0359] 第四替换单元，用于在目标加速算子验证不合格的情况下，对目标加速算子进行分析，根据分析结果调整目标加速算子，然后将第一深度学习模型中的候选算子组合替换为调整后的目标加速算子，并重新执行根据替换后的第一深度学习模型进行推理测试的步骤。

[0360] 基于上述任一实施例，在一个实施例中，第四替换单元，包括：

[0361] 第一替换子单元，用于在目标加速算子是第一加速算子的情况下，根据算子数据处理的先后顺序，逐步融合第一候选算子组合中的算子，或，在目标加速算子是第二加速算子的情况下，根据预设顺序逐步融合第二候选算子组合中的算子；

[0362] 第二替换子单元，用于依次验证融合后的加速算子，确定出导致融合后的加速算子验证不合格的异常算子；

[0363] 第三替换子单元，用于重新构建异常算子，并验证融合重新构建的异常算子后的加速算子，在验证合格的情况下，继续融合下一算子，直至最终融合后的目标加速算子验证合格。

[0364] 基于上述任一实施例，在一个实施例中，装置还包括：

[0365] 加速算子生成模块，用于在未为候选算子组合查找到对应目标加速算子的情况下，针对候选算子组合中的每个算子执行以下处理：在预先建立的基础算子库中查找对应算子的已验证算子；并将与候选算子组合中的多个算子一一对应的多个已验证算子进行融合，得到与候选算子组合所对应的目标加速算子。

[0366] 基于上述任一实施例，在一个实施例中，装置还包括：

[0367] 基础算子生成验证模块，用于创建对应算子的待验证算子，并对待验证算子进行验证，将验证合格的算子确定为算子对应的已验证算子。

[0368] 基于上述任一实施例，在一个实施例中，候选算子确定模块121，包括：

[0369] 算子类型确定单元，用于根据第一深度学习模型的计算图，确定第一深度学习模型中的各个算子的类型；

[0370] 第一算子组合确定单元，用于在第一深度学习模型的计算图的第一分支中，当至少两个相邻的算子满足第一条件、第二条件与第三条件中的任意一项时，将至少两个相邻的算子组成第一深度学习模型的第一候选算子组合；其中，第一条件包括：至少两个相邻的算子均为单射函数算子；第二条件包括：至少两个相邻的算子包括约简算子以及作为约简算子输入的单射函数算子；第三条件包括：至少两个相邻的算子包括能够融合输出的算子与逐元素复用的算子；第一分支为第一深度学习模型的计算图中的任意一个分支；

[0371] 第二算子组合确定单元，用于在第一深度学习模型的计算图的至少两个并行分支中，当存在具有相同上游节点的多个算子时，将多个算子组成第一深度学习模型的第二候选算子组合。

[0372] 基于上述任一实施例，在一个实施例中，装置还包括：

[0373] 样本算子获取模块，用于根据作为样本的第二深度学习模型的计算图，确定第二深度学习模型的样本算子组合；其中，样本算子组合包括第一样本算子组合和/或第二样本算子组合；第一样本算子组合是第二深度学习模型的计算图中，同一分支内的具有依赖关系的多个样本算子的组合；第二样本算子组合是第二深度学习模型的计算图中，并行分支内具有相同上游节点的多个样本算子的组合；

[0374] 基础算子库构建模块，用于获取样本算子组合中的各个样本算子分别对应的已验证样本算子，并保存在基础算子库中；

[0375] 加速算子库构建模块，用于对样本算子组合中的各个样本算子所对应的已验证样本算子进行融合，得到与样本算子组合所对应的加速算子，并保存在加速算子库中。

[0376] 基于上述任一实施例，在一个实施例中，基础算子库构建模块，包括：

[0377] 基础算子创建单元，用于为样本算子组合中的各个样本算子分别创建对应的待验证样本算子；

[0378] 基础算子验证单元，用于对待验证样本算子进行验证；

[0379] 基础算子确定单元，用于验证合格后得到样本算子组合中的各个样本算子所对应的已验证样本算子，并保存在基础算子库中。

[0380] 基于上述任一实施例，在一个实施例中，基础算子验证单元，包括：

[0381] 第一验证子单元，用于从标准深度学习框架中调用与待验证样本算子相对应的对比算子；

[0382] 第二验证子单元，用于为待验证样本算子与对比算子设置相同的数据输入；

[0383] 第三验证子单元，用于获取待验证样本算子对数据输入的第一推理耗时及第一推理结果；

[0384] 第四验证子单元，用于获取对比算子对数据输入的第二推理耗时及第二推理结果；

[0385] 第五验证子单元，用于在第一推理耗时小于第二推理耗时、且第一推理结果与第二推理结果的最大误差小于第一预设阈值的情况下，确定待验证样本算子验证合格。

[0386] 基于上述任一实施例，在一个实施例中，加速算子库构建模块，包括：

[0387] 加速算子创建单元，用于为样本算子组合中的各个样本算子分别调用对应的已验证样本算子，并对已验证样本算子进行融合，得到与样本算子组合所对应的待验证加速算子；

[0388] 加速算子验证单元，用于对待验证加速算子进行验证；

[0389] 加速算子确定单元，用于验证合格后得到与样本算子组合所对应的加速算子。

[0390] 基于上述任一实施例，在一个实施例中，加速算子验证单元，进一步包括：

[0391] 第六验证子单元，用于为样本算子组合与待验证加速算子设置相同的数据输入；

[0392] 第七验证子单元，用于获取待验证加速算子对数据输入的第三推理耗时及第三推理结果；

[0393] 第八验证子单元，用于获取样本算子组合对数据输入的第四推理耗时及第四推理结果；

[0394] 第九验证子单元，用于在第三推理耗时小于第四推理耗时、且第三推理结果与第四推理结果的最大误差小于第二预设阈值的情况下，确定待验证加速算子验证合格。

[0395] 下面对本发明提供的一种机器翻译装置进行描述，下文描述的一种机器翻译装置与上文描述的一种机器翻译方法可相互对应参照。

[0396] 图13是本发明提供的一种机器翻译装置的结构示意图，如图13所示，该装置包括：

[0397] 翻译模块131，用于将待翻译文本输入第一机器翻译模型，得到翻译后的文本；其中，第一机器翻译模型是利用目标加速算子替换初始机器翻译模型中相应的候选算子组合

后得到的；

[0398] 装置还包括：

[0399] 候选算子确定模块1311，用于根据初始机器翻译模型的计算图，确定初始机器翻译模型的候选算子组合；其中，候选算子组合包括第一候选算子组合和/或第二候选算子组合；第一候选算子组合是第一机器翻译模型的计算图中，同一分支内的具有依赖关系的多个算子的组合；第二候选算子组合是第一机器翻译模型的计算图中，并行分支内具有相同上游节点的多个算子的组合；

[0400] 加速算子获取模块1312，用于根据候选算子组合确定对应的目标加速算子；其中，目标加速算子包括第一加速算子和/或第二加速算子；第一加速算子是作为样本的第二机器翻译模型的计算图中，同一分支内的具有依赖关系的多个算子融合后得到的；第二加速算子是第二机器翻译模型的计算图中；并行分支内具有相同上游节点的多个算子融合后得到的；

[0401] 加速算子替换模块1313，用于将目标加速算子替换初始机器翻译模型中的候选算子组合，得到第一机器翻译模型。

[0402] 本实施例中根据第一机器翻译模型的计算图中同一分支内的具有依赖关系的算子或并行分支内具有相同上游节点的算子确定候选算子组合，在算子层面精确、灵活地确定了可融合的算子；根据候选算子组合确定了第一机器翻译模型的加速算子；通过将候选算子组合替换为加速算子，提升了第一机器翻译模型的推理效率，实现灵活高效地对第一机器翻译模型进行推理加速，提升了用户体验，降低了推理设备成本。

[0403] 基于上述任一实施例，在一个实施例中，所述加速算子获取模块1312包括：

[0404] 第一获取单元，用于根据所述候选算子组合在预先建立的加速算子库中查找对应的目标加速算子；其中，所述加速算子库包括算子组合与加速算子之间的映射关系；

[0405] 第二获取单元，用于将所述候选算子组合中的算子进行融合，得到对应的目标加速算子。

[0406] 基于上述任一实施例，在一个实施例中，加速算子替换模块1313，包括：

[0407] 第一替换单元，用于将初始机器翻译模型中的候选算子组合替换为目标加速算子；

[0408] 第二替换单元，用于根据替换后的第一机器翻译模型进行推理测试，根据推理测试的结果对目标加速算子进行验证；

[0409] 第三替换单元，用于在目标加速算子验证合格的情况下，将替换后的第一机器翻译模型确定为第一机器翻译模型；

[0410] 第四替换单元，用于在目标加速算子验证不合格的情况下，对目标加速算子进行分析，根据分析结果调整目标加速算子，然后将初始机器翻译模型中的候选算子组合替换为调整后的目标加速算子，并重新执行根据替换后的初始机器翻译模型进行推理测试的步骤。

[0411] 基于上述任一实施例，在一个实施例中，第四替换单元，包括：

[0412] 第一替换子单元，用于在目标加速算子是第一加速算子的情况下，根据算子数据处理的先后顺序，逐步融合第一候选算子组合中的算子，或，在目标加速算子是第二加速算子的情况下，根据预设顺序逐步融合第二候选算子组合中的算子；

[0413] 第二替换子单元，用于依次验证融合后的加速算子，确定出导致融合后的加速算子验证不合格的异常算子；

[0414] 第三替换子单元，用于重新构建异常算子，并验证融合重新构建的异常算子后的加速算子，在验证合格的情况下，继续融合下一算子，直至最终融合后的目标加速算子验证合格。

[0415] 基于上述任一实施例，在一个实施例中，装置还包括：

[0416] 加速算子生成模块，用于在未为候选算子组合查找到对应目标加速算子的情况下，针对候选算子组合中的每个算子执行以下处理：在预先建立的基础算子库中查找对应算子的已验证算子；并将与候选算子组合中的多个算子一一对应的多个已验证算子进行融合，得到与候选算子组合所对应的目标加速算子。

[0417] 基于上述任一实施例，在一个实施例中，装置还包括：

[0418] 基础算子生成验证模块，用于创建对应算子的待验证算子，并对待验证算子进行验证，将验证合格的算子确定为算子对应的已验证算子。

[0419] 基于上述任一实施例，在一个实施例中，候选算子确定模块1311，包括：

[0420] 算子类型确定单元，用于根据初始器翻译模型的计算图，确定第一机器翻译模型中的各个算子的类型；

[0421] 第一算子组合确定单元，用于在初始机器翻译模型的计算图的第一分支中，当至少两个相邻的算子满足第一条件、第二条件与第三条件中的任意一项时，将至少两个相邻的算子组成初始机器翻译模型的第一候选算子组合；其中，第一条件包括：至少两个相邻的算子均为单射函数算子；第二条件包括：至少两个相邻的算子包括约简算子以及作为约简算子输入的单射函数算子；第三条件包括：至少两个相邻的算子包括能够融合输出的算子与逐元素复用的算子；第一分支为初始机器翻译模型的计算图中的任意一个分支；

[0422] 第二算子组合确定单元，用于在初始机器翻译模型的计算图的至少两个并行分支中，当存在具有相同上游节点的多个算子时，将多个算子组成初始机器翻译模型的第二候选算子组合。

[0423] 基于上述任一实施例，在一个实施例中，装置还包括：

[0424] 样本算子获取模块，用于根据作为样本的第二机器翻译模型的计算图，确定第二机器翻译模型的样本算子组合；其中，样本算子组合包括第一样本算子组合和/或第二样本算子组合；第一样本算子组合是第二机器翻译模型的计算图中，同一分支内的具有依赖关系的多个样本算子的组合；第二样本算子组合是第二机器翻译模型的计算图中，并行分支内具有相同上游节点的多个样本算子的组合；

[0425] 基础算子库构建模块，用于为样本算子组合中的各个样本算子分别创建对应的已验证样本算子，并保存在基础算子库中；

[0426] 加速算子库构建模块，用于对样本算子组合中的各个样本算子所对应的已验证样本算子进行融合，得到与样本算子组合所对应的加速算子，并保存在加速算子库中。

[0427] 基于上述任一实施例，在一个实施例中，基础算子库构建模块，包括：

[0428] 基础算子创建单元，用于获取样本算子组合中的各个样本算子分别对应的待验证样本算子；

[0429] 基础算子验证单元，用于对待验证样本算子进行验证；

[0430] 基础算子确定单元，用于验证合格后得到样本算子组合中的各个样本算子所对应的已验证样本算子，并保存在基础算子库中。

[0431] 基于上述任一实施例，在一个实施例中，基础算子验证单元，包括：

[0432] 第一验证子单元，用于从标准深度学习框架中调用与待验证样本算子相对应的对比算子；

[0433] 第二验证子单元，用于为待验证样本算子与对比算子设置相同的数据输入；

[0434] 第三验证子单元，用于获取待验证样本算子对数据输入的第一推理耗时及第一推理结果；

[0435] 第四验证子单元，用于获取对比算子对数据输入的第二推理耗时及第二推理结果；

[0436] 第五验证子单元，用于在第一推理耗时小于第二推理耗时、且第一推理结果与第二推理结果的最大误差小于第一预设阈值的情况下，确定待验证样本算子验证合格。

[0437] 基于上述任一实施例，在一个实施例中，加速算子库构建模块，包括：

[0438] 加速算子创建单元，用于为样本算子组合中的各个样本算子分别调用对应的已验证样本算子，并对已验证样本算子进行融合，得到与样本算子组合所对应的待验证加速算子；

[0439] 加速算子验证单元，用于对待验证加速算子进行验证；

[0440] 加速算子确定单元，用于通过验证后得到与样本算子组合所对应的加速算子。

[0441] 基于上述任一实施例，在一个实施例中，加速算子验证单元，进一步包括：

[0442] 第六验证子单元，用于为样本算子组合与待验证加速算子设置相同的数据输入；

[0443] 第七验证子单元，用于获取待验证加速算子对数据输入的第三推理耗时及第三推理结果；

[0444] 第八验证子单元，用于获取样本算子组合对数据输入的第四推理耗时及第四推理结果；

[0445] 第九验证子单元，用于在第三推理耗时小于第四推理耗时、且第三推理结果与第四推理结果的最大误差小于第二预设阈值的情况下，确定待验证加速算子验证合格。

[0446] 图14示例了一种电子设备的实体结构示意图，如图14所示，该电子设备可以包括：处理器(processor)1410、通信接口(Communications Interface)1420、存储器(memory)1430和通信总线1440，其中，处理器1410，通信接口1420，存储器1430通过通信总线1440完成相互间的通信。处理器1410可以调用存储器1430中的逻辑指令，以执行上述各提供的一种深度学习模型的推理方法的全部或部分步骤，或，执行上述各提供的一种机器翻译方法的全部或部分步骤。

[0447] 此外，上述的存储器1430中的逻辑指令可以通过软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行本发明各个实施例方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(ROM,Read-Only Memory)、随机存取存储器(RAM,Random Access Memory)、磁碟或者光盘等各种可以

存储程序代码的介质。

[0448] 另一方面，本发明还提供一种计算机程序产品，计算机程序产品包括存储在计算机可读存储介质上的计算机程序，计算机程序包括程序指令，当程序指令被计算机执行时，计算机能够执行上述各提供的一种深度学习模型的推理方法的全部或部分步骤，或，执行上述各提供的一种机器翻译方法的全部或部分步骤。

[0449] 又一方面，本发明还提供一种计算机可读存储介质，其上存储有计算机程序，该计算机程序被处理器执行时实现以执行上述各提供的一种深度学习模型的推理方法的全部或部分步骤，或，执行上述各提供的一种机器翻译方法的全部或部分步骤。

[0450] 以上所描述的装置实施例仅仅是示意性的，其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性的劳动的情况下，即可以理解并实施。

[0451] 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到各实施方式可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件。基于这样的理解，上述技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品可以存储在计算机可读存储介质中，如ROM/RAM、磁碟、光盘等，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行各个实施例或者实施例的某些部分所述的方法。

[0452] 最后应说明的是：以上实施例仅用以说明本发明的技术方案，而非对其限制；尽管参照前述实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

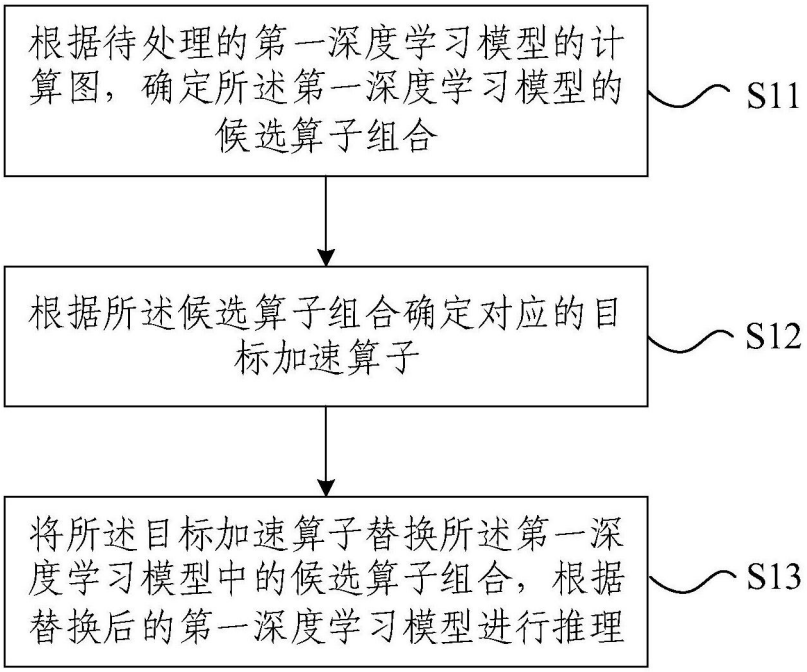

根据待处理的第一深度学习模型的计算图，确定所述第一深度学习模型的候选算子组合
S11
根据所述候选算子组合确定对应的目标加速算子
S12
将所述目标加速算子替换所述第一深度学习模型中的候选算子组合，根据替换后的第一深度学习模型进行推理
S13

图1

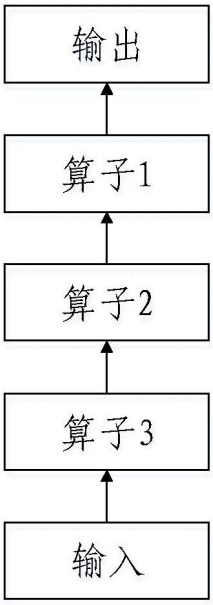

输出
算子1
算子2
算子3
输入

图2

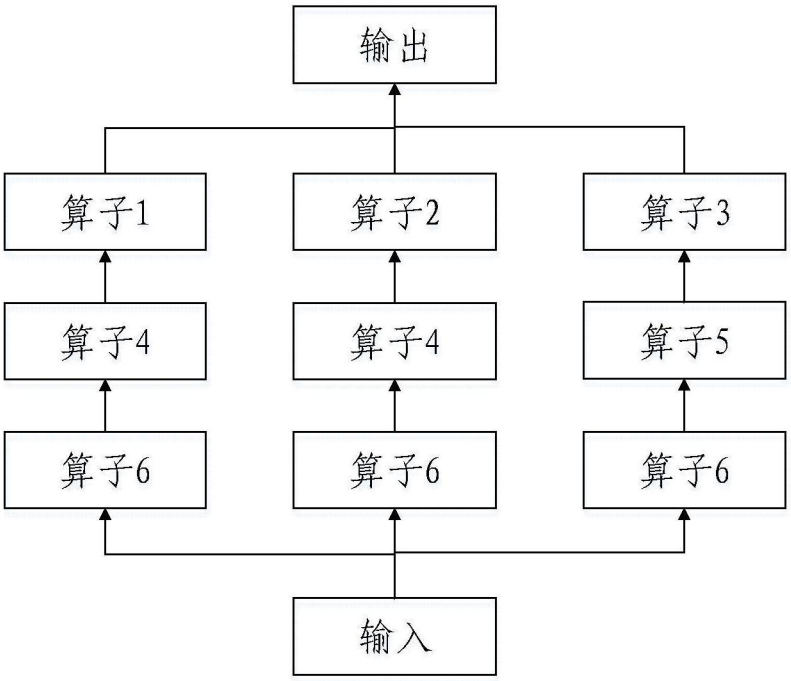


图3

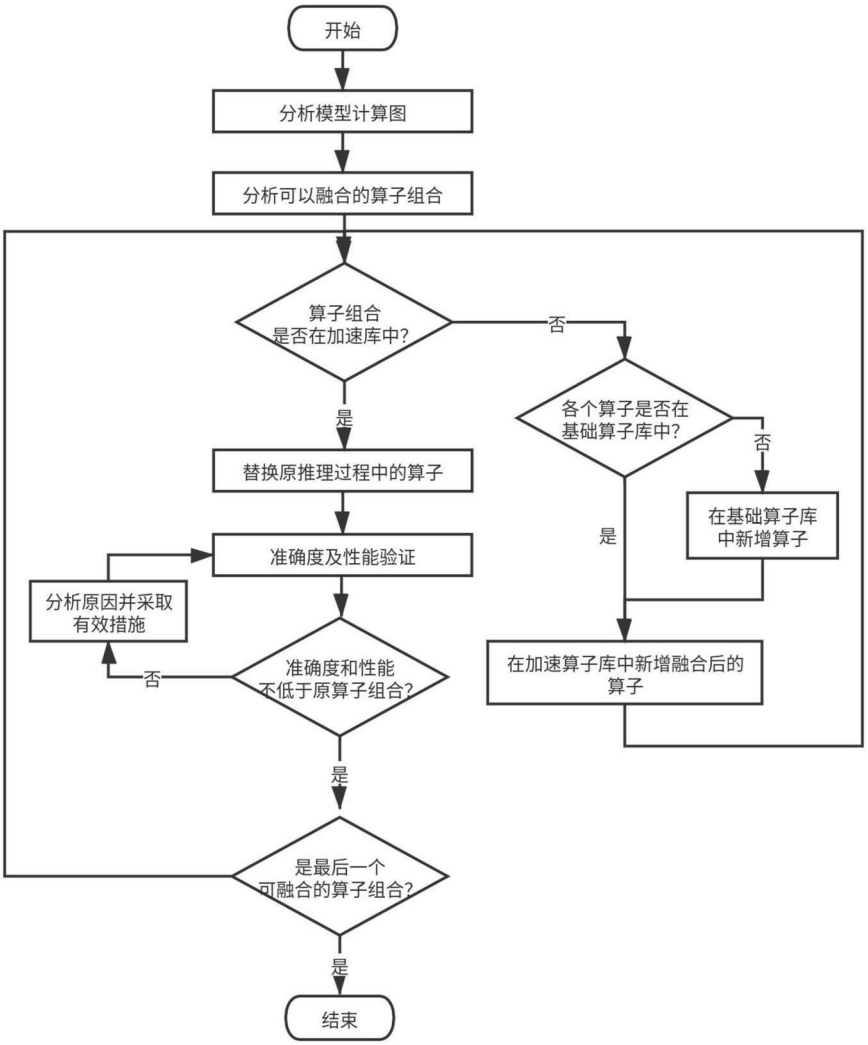

开始
分析模型计算图
分析可以融合的算子组合
算子组合
是否在加速库中?
否
是
各个算子是否在
基础算子库中?
否
是
替换原推理过程中的算子
在基础算子库
中新增算子
准确度及性能验证
分析原因并采取
有效措施
准确度和性能
不低于原算子组合?
否
在加速算子库中新增融合后的
算子
是
是最后一个
可融合的算子组合?
是
结束

图4

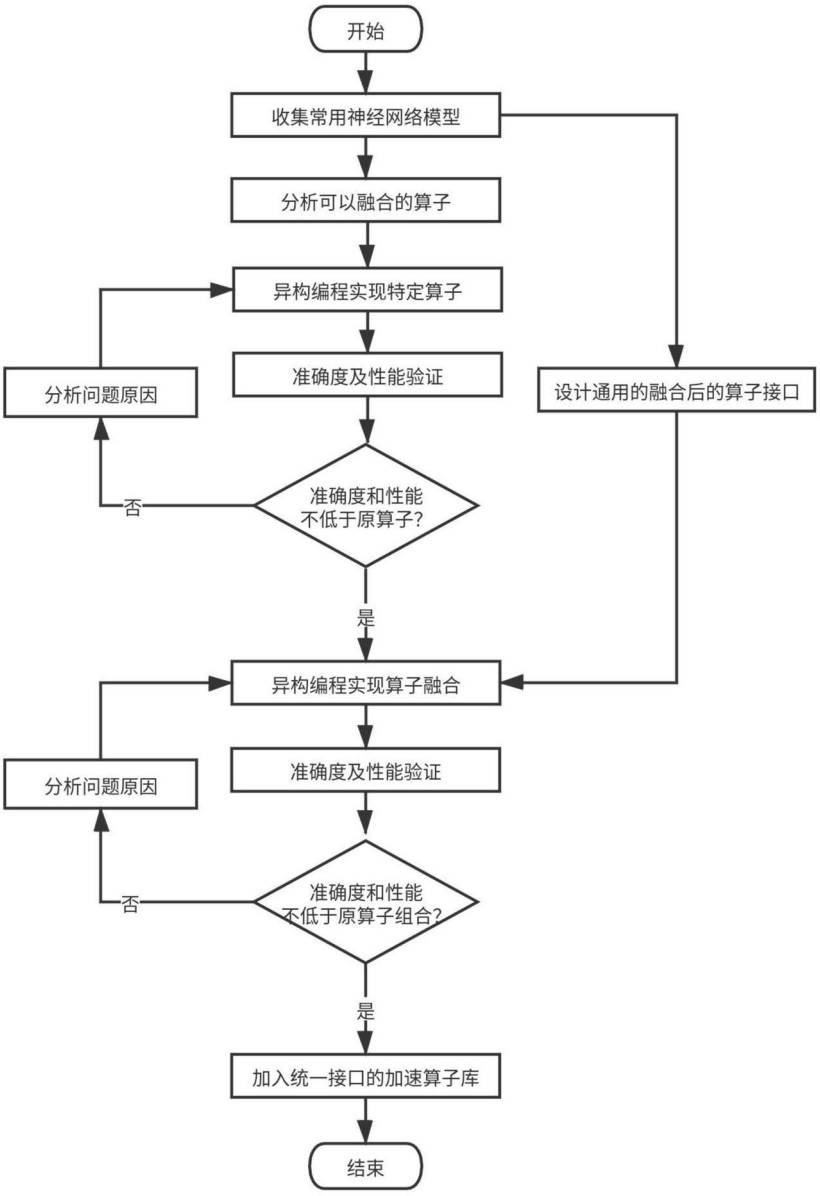

开始
收集常用神经网络模型
分析可以融合的算子
异构编程实现特定算子
准确度及性能验证
分析问题原因
设计通用的融合后的算子接口
准确度和性能
不低于原算子?
否
是
异构编程实现算子融合
准确度及性能验证
分析问题原因
准确度和性能
不低于原算子组合?
否
是
加入统一接口的加速算子库
结束

图5

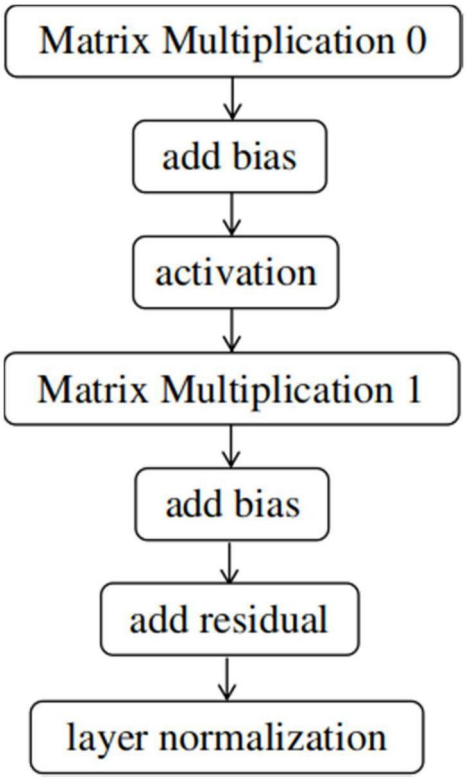

Matrix Multiplication 0
add bias
activation
Matrix Multiplication 1
add bias
add residual
layer normalization

图6

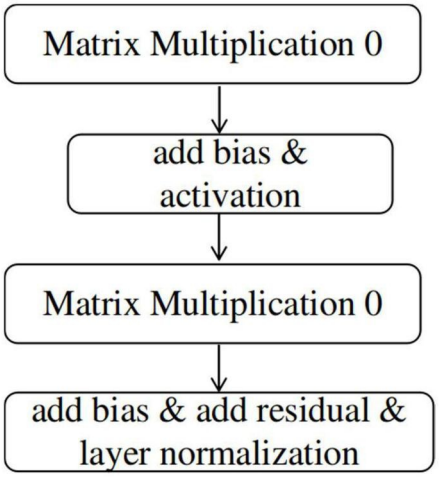

Matrix Multiplication 0
add bias &
activation
Matrix Multiplication 0
add bias & add residual &
layer normalization

图7

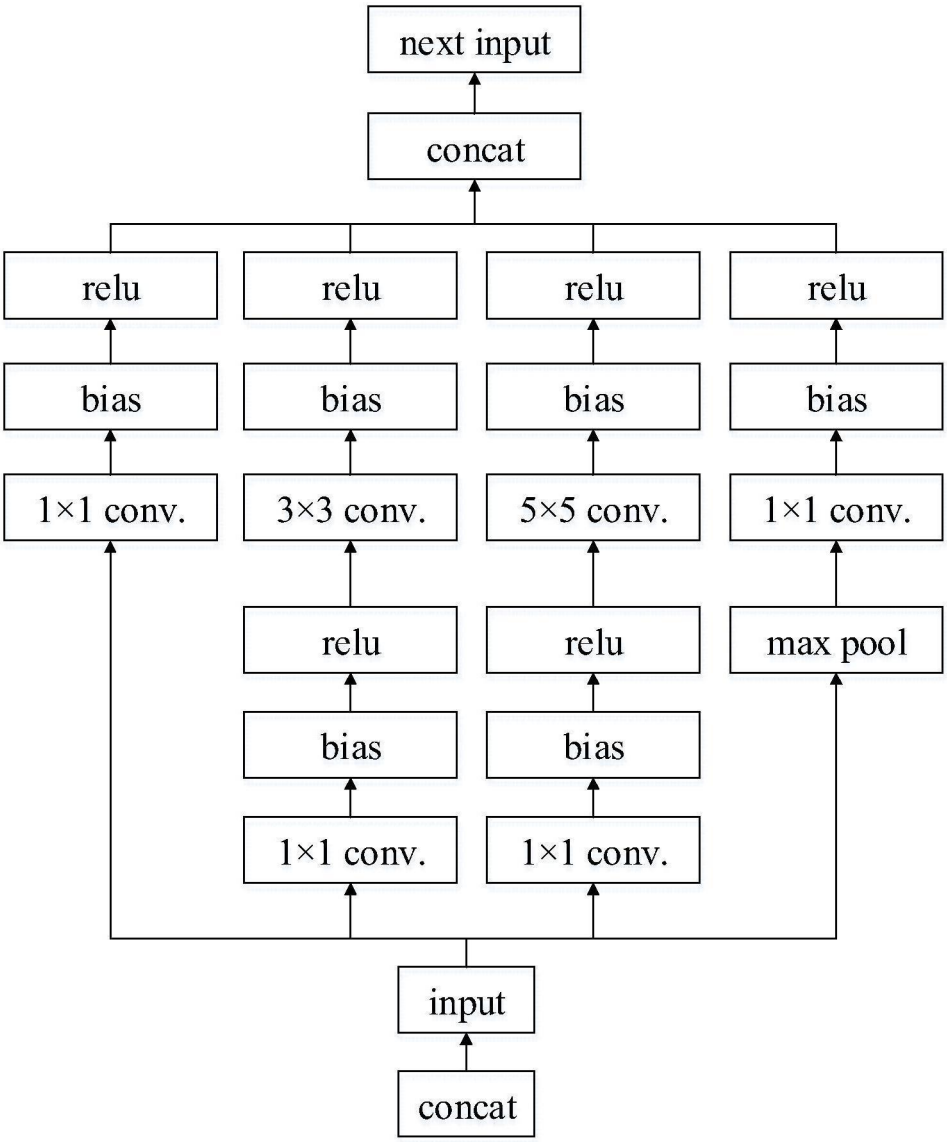

next input
concat
relu
relu
relu
relu
bias
bias
bias
bias
1×1 conv.
3×3 conv.
5×5 conv.
1×1 conv.
relu
relu
max pool
bias
bias
1×1 conv.
1×1 conv.
input
concat

图8

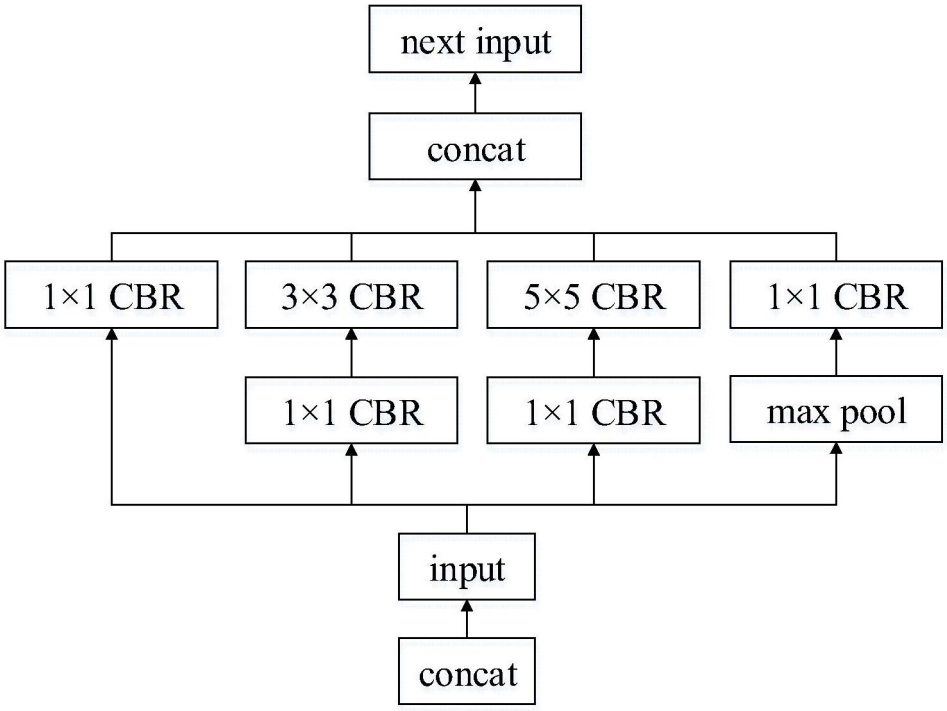

next input
concat
1×1 CBR
3×3 CBR
5×5 CBR
1×1 CBR
1×1 CBR
1×1 CBR
max pool
input
concat

图9

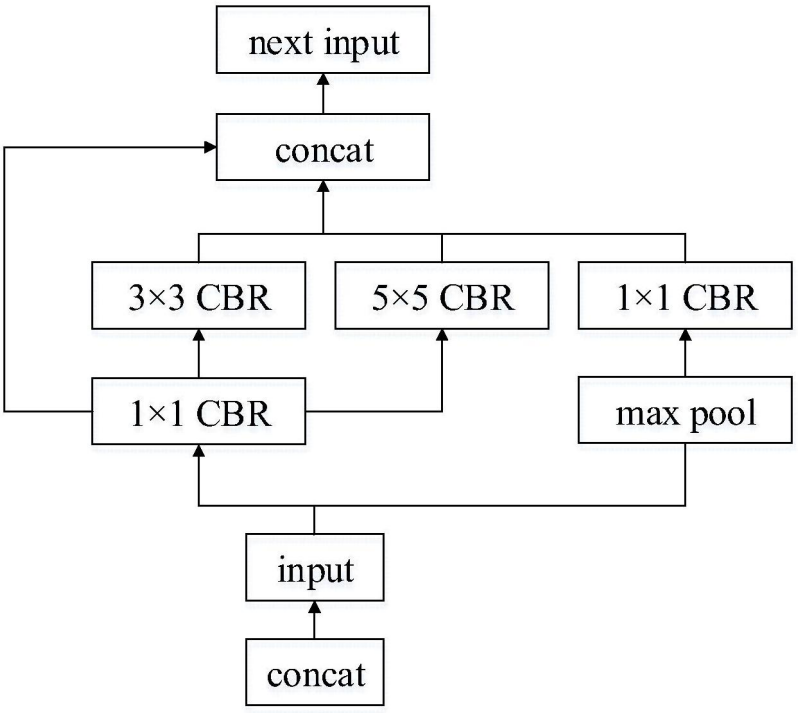

next input
concat
3×3 CBR
5×5 CBR
1×1 CBR
1×1 CBR
max pool
input
concat

图10

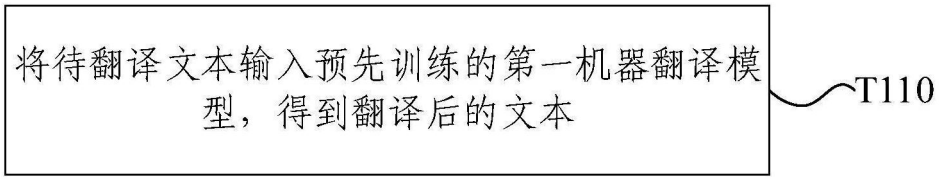

将待翻译文本输入预先训练的第一机器翻译模型，得到翻译后的文本
T110

图11

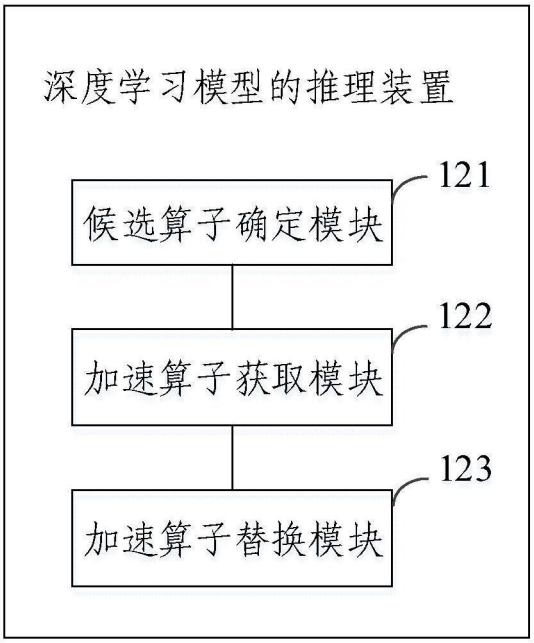

深度学习模型的推理装置
候选算子确定模块
121
加速算子获取模块
122
加速算子替换模块
123

图12

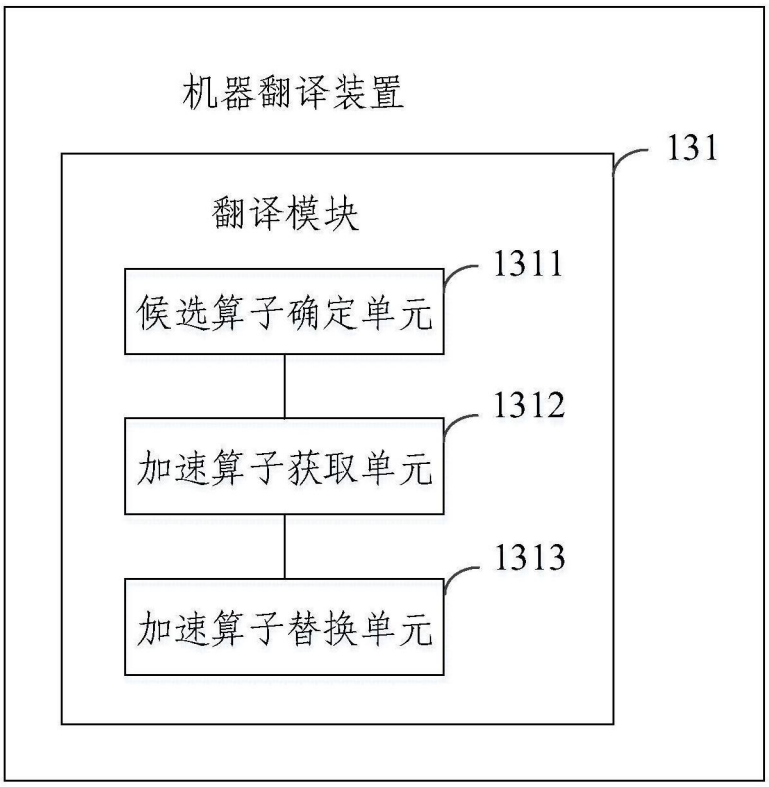

机器翻译装置
131
翻译模块
1311
候选算子确定单元
1312
加速算子获取单元
1313
加速算子替换单元

图13

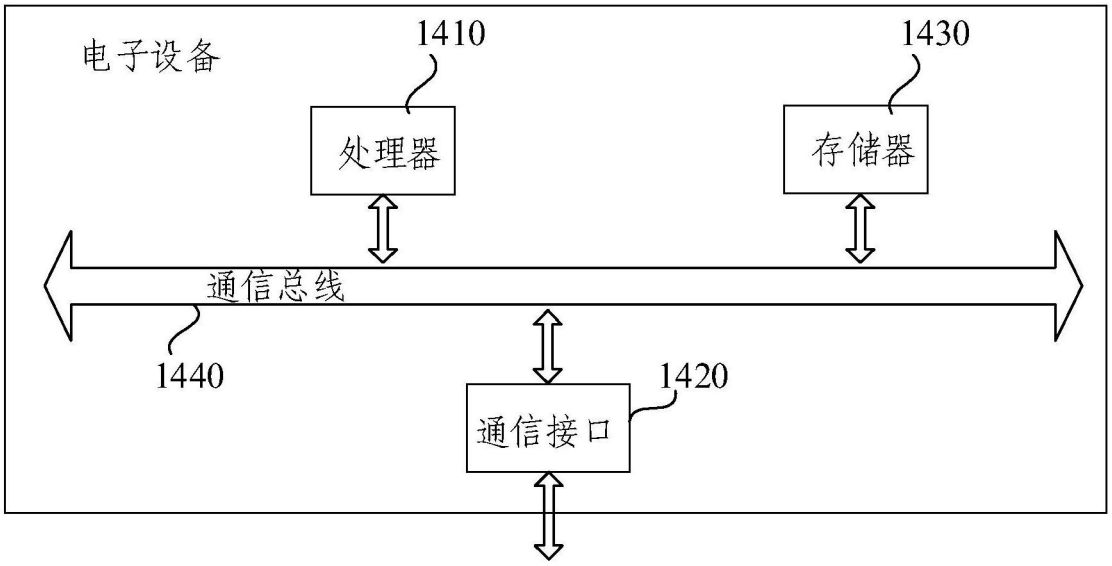

电子设备
1410
处理器
1430
存储器
通信总线
1440
1420
通信接口

图14